Population genetic studies of the S-locus gene family and other loci in self-compatible and self-incompatible populations of the plant *Antirrhinum*

Cristina Alexandra Gonçalves Paula Vieira

Thesis presented for the degree of Doctor of Philosophy at the University of Edinburgh

2000

Abstract

The study of genetic variation within and among closely related species is a central concern of population genetics. Such studies can provide information on various evolutionary forces that shape variation, such as mutation, selection, migration, genetic drift, recombination, mating system differences, and effective population sizes. In plants, the mating pattern varies considerably and ranges from regular systems of inbreeding by self fertilisation to strict outcrossing maintained by incompatibility systems. In highly inbred populations there is an increased frequency of homozygotes, resulting in reduced effective population size, and lowered effective rates of genetic recombination. Both adaptive and purifying selection in regions of effective low recombination are known to result in low levels of variability at linked neutral sites. Bottlenecks may also be more extreme in inbreeders than in outcrossing species. These predictions have been tested by comparing inbred and outbreeding species, but there are few data on diversity within populations. However, total species diversity levels are expected to be less reduced in inbreeders than within-population values.

In genera such as *Antirrhinum*, closely related species have different breeding systems, but other factors that affect the levels and patterns of genetic diversity are not expected to vary greatly. Therefore studies within such genera should be ideal to test for effects of the mating system. In this work the mating system of several populations and species of *Antirrhinum* were established in the glasshouse. Levels of DNA diversity were estimated based on *cyc* and *fil1* nuclear genes. Both genes are shown to belong to gene families. In these gene families, some members are very

similar, which makes it difficult to determine orthology. In the cases where orthology is not a problem, low levels of nucleotide diversity were found. Therefore the effect of the mating system on genetic diversity could not be tested. I found unexpectedly little divergence between several *Antirrhinum* species, *Digitalis*, and the more distantly related genus *Verbascum* for genes of the *cyc* and *fil1* gene families. The generality of this pattern was addressed by extending these studies to *fil2*, *far*, *globosa* and *Adh* genes. Evidence is shown that these genes are also members of gene families in *Antirrhinum*. For *fil2*, *far*, and *globosa*, very similar sequences were found in *Antirrhinum* and *Verbascum*. For *Adh* I could not determine orthology because repeated gene duplication and loss of elements in this gene family has occurred in the *Antirrhinum* and *Verbascum* lineages. Several hypotheses that could account for the low diversity and divergence are discussed.

In *Antirrhinum*, self-incompatibility is controlled by a gametophytic system. The gene responsible for pistil self-incompatibility is the S-locus that encodes basic glycoproteins with ribonuclease activity. High levels of variability are observed, consistent with frequency-dependent selection. The putative targets of selection are those regions, such as the hypervariable regions of this gene, that may be involved in specificity determination. In order to gather evidence on whether these regions are hypervariable because they are the target of selection, or merely regions of relaxed selective constraint, I have partially sequenced *Antirrhinum* S-alleles and analysed their level and pattern of nucleotide diversity. Within each allelic type, low levels of diversity were observed. Similar alleles were found in self-compatible and self-incompatible species, suggesting that the *Antirrhinum* group evolved recently.

Acknowledgements

I would like to express my deepest gratitude to my supervisor and good friend. She helped me mature and develop as both a scientist and person over the last 3 years.

Special thanks go Dr Jorge Vieira and all my colleagues at the Charlesworth Laboratory. The many scientific discussions we had, much improved the final version of this work.

There are two friends who deserve special recognition, Professor Daniel Hartl and Professor Claudio Sunkel. They were the ones who encourage me to follow my dream of becoming a scientist.

Special thanks go to the Commission of European Communities for the grant (Grant ERBFMBICT 972455), that made possible the realisation of this work.

Many thanks to everyone involved in the field-work, especially Dr. Carlos Aguiar, Eng. Ana Carvalho, my sister and my parents.

Mostly I wish to extend my never-ending gratitude to my family, specially my husband, for their constant support, their love, and their commitment to help me achieve the goal of being the best I could ever be.

Table of Contents

Chapter 3

The breeding system and variation in four morphometric characters of Antirrhinum *and* Misopates *populations.*

Chapter 4

Evolution of the cycloidea gene family in Antirrhinum *and* Misopates.

Chapter 5

Low diversity and divergence in the fil1 gene family of Antirrhinum.

Chapter 6

Low rates of silent substitution in nuclear genes of two distantly related Scrophulariaceae (Antirrhinum *and* Verbascum*).*

Chapter 1 - Introduction

A review of the effect of the breeding system on levels of nucleotide variation and population genetic studies of the gametophytic S-locus.

1.1. Introduction

Understanding the genetic and ecological mechanisms that shape intraspecific DNA sequence polymorphism and divergence in natural populations is a fundamental goal of evolutionary biology research. The level and pattern of nucleotide polymorphisms, both within and between species, can give information about population structure and the action of natural selection on DNA and protein sequences. The latter is revealed as deviations from the level and patterns expected if the variation is selectively neutral (Kimura 1983; Ohta 1992; Kreitman and Akashi 1995). Under the neutral or nearly neutral theory of evolution (our null hypothesis), the ultimate fate of any nucleotide variant is to become fixed or lost (Kimura 1983; Ohta 1992). Selection may lead to the fixation of a favourable allele, but depending on the nature of selection, a stable polymorphism can also be maintained in a population (see for instance Kreitman and Hudson 1991). Selection affects the frequency distribution of variants around the

site under selection, in a fashion which depends on the type of selection and its strength, as well as the amount of recombination in the region of interest. Several test statistics have been developed (see for instance, Hudson *et al.* 1987; Tajima 1989; McDonald and Kreitman 1991; Fu and Li 1993; Hudson *et al.* 1994; Kelly 1997; Wall 1999), which can be used both to detect significant deviations from neutral expectation and to infer the nature of selection. This approach is very powerful, because it can lead to the identification of the region, or in some cases even the amino acid or nucleotide site, that is under selection and differences in fitness (as low as on the order of the inverse of the population size; Kimura 1983) among genotypes can be estimated. Also, by understanding how a gene, both within and between species, is responding to different selection pressures, a better understanding of gene evolution can be obtained. However, it is unlikely that from these analyses alone we will be able to identify the feature that is under selection.

Most of the information relevant to this area has been obtained for *Drosophila* populations (see Powell 1999). The ecology and genetics of plant species are, however, different from those of *Drosophila* in many respects. One main difference is the mating system. In plant species it varies from predominantly selfing to complete outcrossing (Stebbins 1957; Barrett and Eckert 1990) and it is a major determinant of the distribution of genetic variation within and between plant populations (Shoen and Brown 1991). It is becoming clear that considerable differences in sequence diversity exist in plants with different breeding systems. However there are as yet few within population sequencing data from plant species (Liu *et al.* 1998, 1999; Savolainen *et al.* 2000). Additional data are therefore needed in order to test the effects of breeding system on nucleotide

diversity. Furthermore, comparing levels of neutral variability with variability in the S-locus (the gene known to control self-incompability and to be under balancing selection), together with knowledge on the specificity of the alleles being studied, may help elucidate what are the regions involved in determining specificity.

1.2. The effect of the breeding system on levels of variability

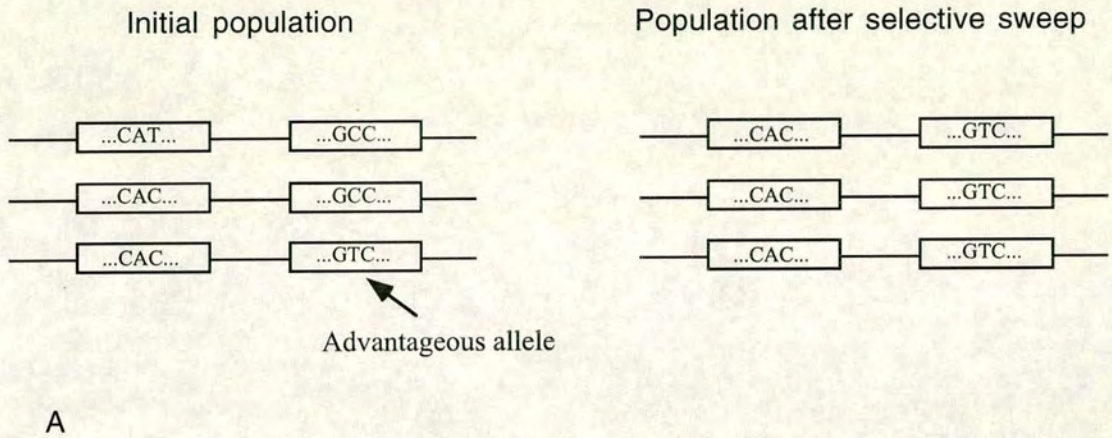
1.2.1. Theoretical expectations

Under the classical theory of neutral molecular evolution, allelic diversity depends on the effective population size (N_e) and the mutation rate (μ; Kimura 1983). In a highly inbred population each individual carries two identical genomes, and the population thus behaves similarly to one of half its actual size and it is expected to have half of the neutral diversity levels of a similar outcrossing population (Pollak 1987). Therefore within populations a halving of variability levels is expected in the absence of other forces (Charlesworth *et al.* 1993). Further decreases in variability levels are expected since bottlenecks may be more extreme in inbreeders, in which a single seed can found a new population, in contrast with outcrossing species (Schoen and Brown 1991). Many selfing species may also be more likely than outcrossers to occur in a metapopulation system, one in which high rates of extinction will lead to lower effective population size and thus reduced variability (Wade and McCauley 1988; Barton and Whitlock 1997). In addition, polymorphisms maintained by heterozygote advantage will tend to be lost in populations that are highly inbreeding (Kimura and Ohta 1971; Charlesworth and Charlesworth 1995). Moreover highly inbreeding populations have lower effective rates

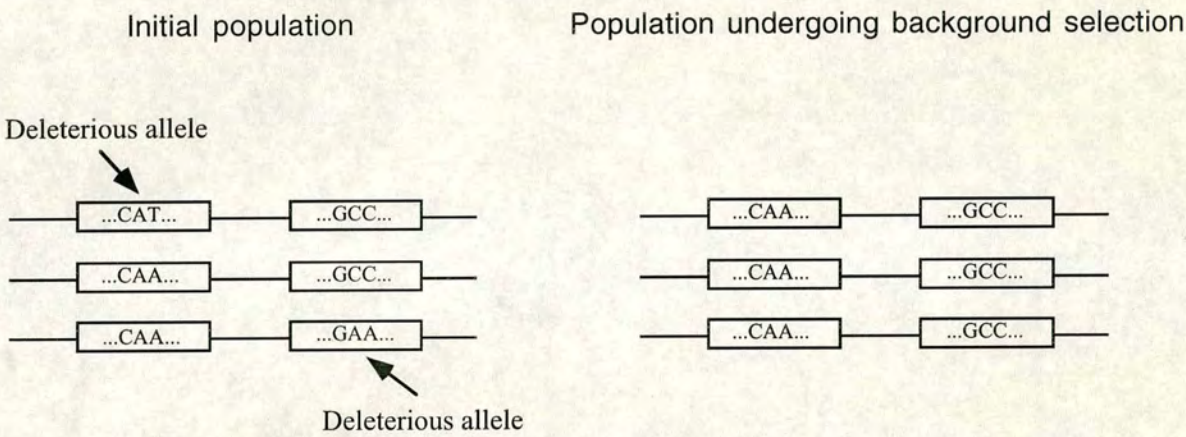
of genetic recombination. As a consequence, as a favourable mutation sweeps through a population, linked neutral variants will "hitch-hike" along to high frequency, reducing variability levels (Maynard Smith and Haigh 1974; Fig 1.1A). Also, as purifying selection eliminates slightly deleterious alleles, that constantly arise through mutation, from a population, linked neutral variants are eliminated as well, reducing variability levels (Charlesworth *et al.* 1993; Fig. 1.1B). Indeed in *Drosophila melanogaster,* genes localised in regions of reduced rates of meiotic crossing-over (such as the telomeric region of the X chromosome, the base of all chromosomal arms and the entire chromosome 4) are less polymorphic than genes in regions of normal crossing-over rates, but there is no correlation between recombination and the level of inter-species divergence (Aguadé *et al.* 1989; Begun and Aquadro 1992, Aquadro *et al.* 1994; Moriyama and Powell 1996). A similar correlation between DNA variation and local recombination rates have been found in several other organisms (reviewed by Charlesworth and Charlesworth 1998) including the plants *Lycopersicon* (Stephan and Langley 1998), *Aegilops* (Dvorak *et al.* 1998) and *Beta vulgaris* subsp. *maritima* (Kraft *et al.* 1998).

1.2.2. Empirical data

When testing the effect of the breeding system on levels of variability it is important to consider the sampling strategy used. Most studies involve samples taken from different populations of a species and thus represent species-level diversity not within population diversity. Species-level diversity is insensitive to the effect of the breeding system compared with within population diversity (Charlesworth *et al.* 1997). In highly inbred species, if within-species diversity is high but within population



A



B

Fig. 1.1. Selective sweeps (A) and background selection (B) in a population of chromosomes. (A) The new advantageous allele will go to fixation. In the absence of recombination linked neutral sites will be fixed as well. (B) Deleterious alleles that constantly arise through mutation in the population will be eliminated. In the absence of recombination linked neutral sites will be eliminated as well. Therefore, both processes result in reduced neutral variation in regions of reduced crossing over.

variability is low (see 1.2.1.), a large number of differences must be fixed between populations (Charlesworth and Pannell 2000).

As theory predicts, selfing in plants is correlated with reduced within-population allozyme variability (Brown 1979; Rick *et al.* 1977; Barrett and Husband 1990; Hamrick and Godt 1990, Schoen and Brown 1991; Fenster and Ritland 1992; Hamrick and Godt 1996; Charlesworth and Yang 1998). Overall the data indicate that allozyme diversity in selfing plant populations is ~50% of that of obligate outcrossers (Hamrick and Godt 1990; Schoen and Brown 1991). In the absence of any other factor, theory predicts that in highly inbred populations levels of variability should be halved compared with outcrossing populations. However, these comparisons involve distantly related species and factors other than the breeding system may differ causing difficulties in the interpretation of the results. Indeed when comparisons within genera are performed larger reductions in variability levels are observed (Rick *et al.* 1977; Barrett and Husband 1990, Fenster and Ritland 1992; Charlesworth and Yang 1998). Furthermore, it is possible that some of the amino acid variation is selectively maintained, again causing difficulties in the interpretation of the results.

RFLP data can also be used to estimate nucleotide diversity, given detailed information on the nature of the variants (see Stephan and Langley 1998). Dvorak *et al.* (1998) analysed levels of variability for 52 single copy RFLP loci in five self-compatible and one self-incompatible *Aegilops* species. These authors always found greater than twofold higher levels of variability for the self-incompatible than for the self-compatible species. Stephan and Langley (1998) analysed levels of variability for 36 RFLP loci across the genomes of eight *Lycopersicon* species with different mating systems. These

authors also always found a difference greater than twofold between selfing species and self-incompatible species. These effects argue for hitchhiking, background selection, higher extinction rates or bottleneck effects in inbred species. However RFLP data are based on the number of different band lengths detected on gels and new allelic variants can arise not only by mutation but also by recombination. Since self-compatible species have lower effective recombination rates, this may result in an overestimation of the diversity in outcrossers (Charlesworth and Pannell 2000).

Microsatellite diversity is often estimated using the variance in the number of repeats, that take into account the special features of the way these variants are thought to arise (Estoup and Cornuet 1999). Furthermore, variability levels depend on microsatellite structure. Therefore comparing different sets of microsatellite markers requires caution (Treuren *et al.* 1997). Nevertheless in *A. thaliana* (a highly selfing species with less than 1% of outcrossing; Abbott and Gomes 1988) no microsatellite polymorphism was deteced within any of three within-population samples, although overall allelic diversity was very high for 12 Japanese populations (Todokoro *et al.* 1995), but *A. lyrata* (an outbreeder) populations were quite variable (Treuren *et al.* 1997). Based on a set of six microsatellite loci, variability levels in *Mimulus laciniatus* populations (selfing rates ranging from 0.72 to 1.00) are more than twofold lower than in *Mimulus guttatus* populations (selfing rates ranging from 0.46 to 0.55; Awadalla and Ritland 1997). This pattern was also observed for three out of four microsatellite loci that could be compared between *M. guttatus* (outcrossing) and *M. nasutus* (self-fertilising; Kelly and Willis 1998).

DNA sequence analyses can provide data on silent variants (synonymous, intron and possibly 5' and 3' flanking regions), that are probably close to neutral. When levels of diversity of alcohol dehydrogenase (*Adh*) from different species (*Zea mays*, *Leavenworthia stylosa*, *Pennisetum glaucum*, *Arabidopsis thaliana* and *Hordeum vulgare*) are compared, lower levels of diversity are observed in self-fertilising than in cross-fertilising species, the outbreeding species *Pennisetum glaucum* being the exception (Cummings and Clegg 1998; Charlesworth and Charlesworth 1998). The low level of diversity in the latter could be attributed to *Adh* in this species being located in a region of low recombination. However, this comparison suffers from the same difficulty as some of the allozyme studies, namely distantly related species are being compared.

Few DNA sequencing studies comparing nucleotide diversity in closely related selfing and outcrossing species are available (in *Leavenworthia* - Liu *et al.* 1998, 1999; and *Arabidopsis* - Savolainen *et al.* 2000). Levels of genetic diversity at the *Adh* and phosphoglucose isomerase (*PgiC*) loci within and between populations were compared between outcrossing and inbreeding populations in the annual plant genus *Leavenworthia* (Liu *et al.* 1998; 1999). For *Adh* within inbred populations of *L.uniflora* and *L.crassa*, no DNA sequence variants were seen among the alleles sampled. For *PgiC* the highly selfing species *L. uniflora* and *L.torulosa* have also no within-population variation. However, high diversity was seen in alleles of the self-incompatible *L. crassa* populations for both loci. For *Adh* the outcrosser *L. stylosa* also has high levels of variability. The *PgiC* locus of *L. stylosa* seems to be under balancing selection and therefore is not suitable for such comparisons (Liu *et al.* 1999; Filatov and Charlesworth 1999).

In *Arabidopsis,* nucleotide variation at the *Adh* locus was studied in three populations of *A. lyrata* ssp. *lyrata*/ ssp. *petraea* (outcrossing species) and compared with nucleotide polymorphism of *A. thaliana* (Savolainen *et al.* 2000). The variation within *A. thaliana* populations is reduced more than half compared with the outcrossing species.

Therefore, since the effect of inbreeding appears to be generally greater than a twofold reduction in overall diversity, some process such as selection for advantageous mutations, or against deleterious mutations, or bottlenecks occurring predominantly in inbreeders appear to be necessary to account for these findings.

1.3. The effect of the breeding system on F_{ST}

It is well known that divergence between populations of selfing plant species is consistently higher (as measured by F_{ST}) than divergence between populations of self-incompatible species. It has been estimated that between population allozyme divergence values of selfing plants are up to five times greater than those of outcrossers (Hamrick and Godt 1996). This pattern is also observed for nucleotide data (Liu *et al.* 1998; 1999; Savolainen *et al.* 2000). Much of the difference is caused by the approximately twofold higher within-population diversity in outbreeding species (see above). However neutral models show that selfing alone should only increase F_{ST} values by at most a factor of two (Maruyama and Tachida 1992; Nordborg 1997; Charlesworth *et al.* 1997). Therefore the very high F_{ST} values observed between selfing populations must be attributable to additional factors. Both background selection, local selection, and bottlenecks associated with colonisation events could potentially account for these high F_{ST} values

(Charlesworth *et al.* 1997). However it remains difficult to distinguish between the various possibilities (Liu *et al.* 1999).

1.4. The breeding system

Fertilisation in flowering plants begins with a pollen grain bearing the male gametes landing on the female stigma. However in some angiosperms, if the pollen grain lands on its own stigma or on the stigma of genetically related plants, fertilisation will fail either because the pollen grain fails to germinate, or the pollen tube does not reach the ovules (de Nettancourt 1977). This is known as self-incompatibility. There are several ways in which the genetic barrier between pollination and fertilisation in angiosperms is achieved (Bateman 1952; Lewis 1954; Pandey 1957; Lundqvist 1975; Arasu 1968).

There are two types of self-incompatibility (SI): homomorphic and heteromorphic. In the homomorphic, the simplest, best studied and most common SI system, occurring in as many as 60- 90 angiosperm families (East 1940, Charlesworth 1985; Gibbs 1988), no polymorphism in floral morphology is associated with SI. In the heteromorphic type, flowers of the same species can have two (distyly) or three (tristyly) floral morphological types and pollination is compatible only between different morphological types (de Nettancourt 1977).

The homomorphic type is further classified into gametophytic (GSI) and sporophytic (SSI), based on whether the pollen behaviour in the SI reactions is determined by its own haploid genotype or by the diploid genotype of its parental plant (Fig. 1.2). In GSI the incompatibility reaction of the pollen grain is determined by its own genotype, therefore identity between the allele occurring in the pollen grain and one

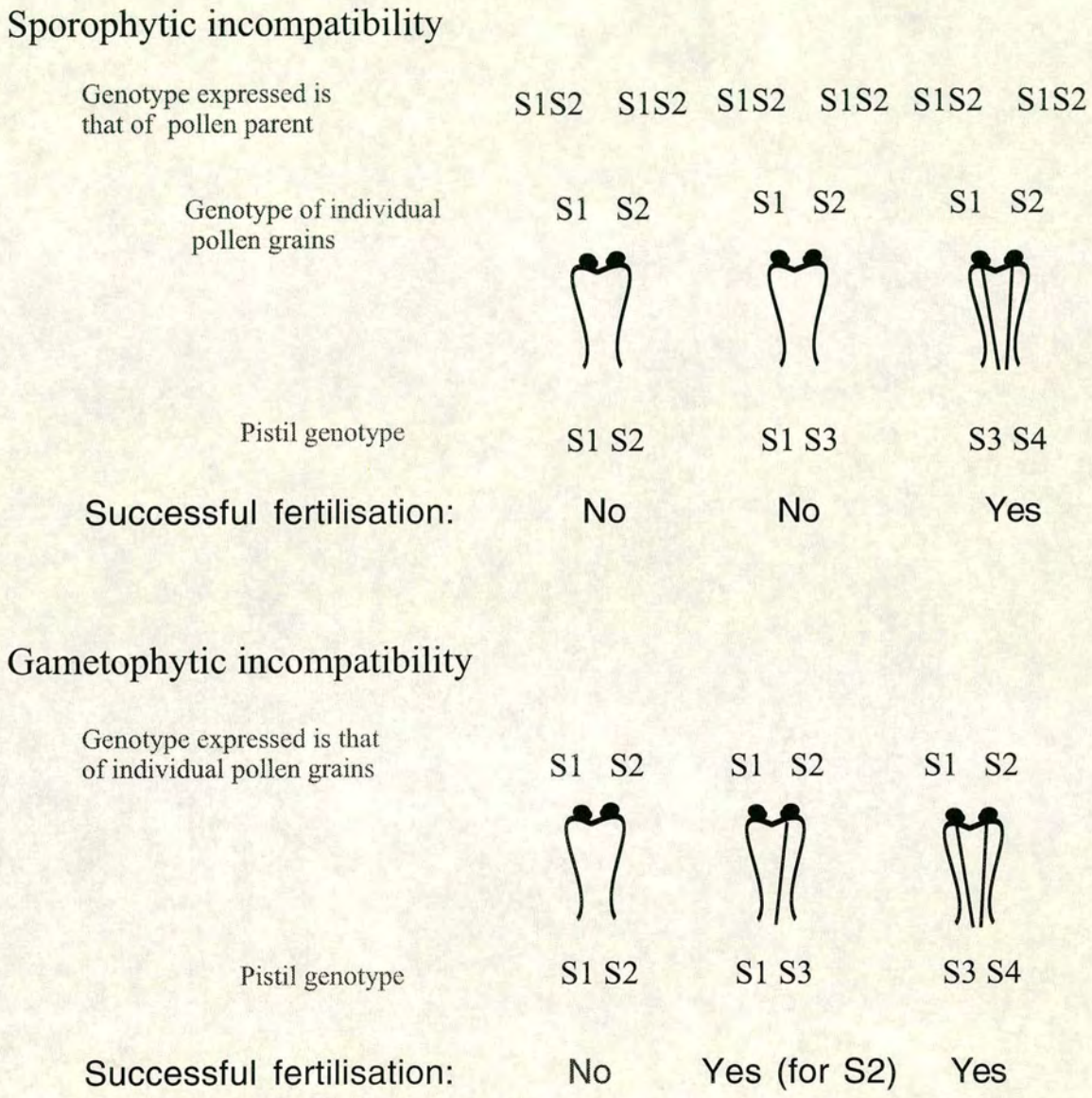


Fig. 1.2. Behaviour of pollen in the two major self-incompatibility systems (modified from Anderson *et al.* 1983).

of the alleles in the pistil is sufficient to prevent fertilization. The number of loci controlling GSI is variable, ranging from one to four, although the number of cases with more than a single locus is small (Lundqvist 1975). For instance in grasses where SI is under the control of two unlinked multiallelic genes S and Z, self-fertilisation is prevented when both S and Z alleles present in the pollen are matched in the style (Hayman 1956; Hayman and Richter 1992). Also in the families Ranunculaceae and Chenopodiaceae complementary S gene systems involving at least three or four S loci have been revealed (Lundqvist 1975). SSI involves determination of the pollen incompatibility reaction by the parental sporophyte. All pollen grains produced by a plant have therefore the same specificity even though they may have different genotypes. Nearly all studies have shown that this system is governed by a single locus (de Nettancourt 1977). These two systems usually differ in several other characteristics such as: the place where the SI response generally takes place (within the transmitting tissue of the style in GSI, at the stigma surface in SSI); the recognition specificities of the S-alleles (codominance in the majority of the cases but complex dominant relationships can occur in SSI); the viability of stored pollen (long-term viability in GSI but short in SSI); the presence of wet (GSI) or dry stigmas (SSI); the presence of binuclear (GSI) or trinuclear (SSI) pollen (with two exceptions out of 23 flowering plant families analysed, SSI is associated with trinuclear and GSI with binuclear pollen; see Brewbaker 1957; 1967; Sims 1993; Weller *et al.* 1995). Molecular studies have been performed in Solanaceae, Papaveraceae, Rosaceae, Scrophulariaceae (GSI) and in Brassicaceae (SSI).

1.4.1. Evidence that S proteins control recognition and rejection of self-pollen by the pistil

Evidence that S-proteins control recognition and rejection of self-pollen was obtained in the last 20 years using a variety of approaches. First, pistil-specific proteins are found in high concentrations in the transmitting tract of the style (the site at which inhibition of pollen tubes occurs during incompatible matings; Anderson *et al.* 1986; Cornish *et al.* 1987), at the time of transition of self-compatible immature buds to SI mature flowers (Clark *et al.* 1990; Kheyr-Pour *et al.* 1990). These different pistil specific proteins, called S-proteins, have different electrophoretic mobilities and co-segregate in families with different incompatibility types. Kovaleva and co-workers demonstrated a correlation between these proteins and an increase in style ribonuclease activity (Kovaleva and Musatova 1975; Kovaleva *et al.* 1978; Kovaleva 1983). Kamboj and Jackson (1986) further found that these proteins have molecular weights in the region of 24,000 to 33,000, and that they have stretches of basic residues. However a very significant contribution was made by Mau *et al.* (1986) when they reported the 15 N-terminal amino acids of these proteins and showed that these are highly conserved among different S-proteins. Several groups took advantage of this finding to design approaches for cloning S-allele cDNAs. Since then, molecular studies in Solanaceae, Scrophulariaceae and Rosaceae have revealed that the S-locus proteins in the pistil are a class of extracellular basic glycoproteins with ribonuclease activity, called S-RNases (Newbigin *et al.* 1993; Matton *et al.* 1994; Broothaerts *et al.* 1995; Golz *et al.* 1995; Kao and McCubbin 1996; Sassa *et al.* 1996; Xue *et al.* 1996) as predicted by the earlier work. As expected for self-incompatibility molecules, the sequences of the S-proteins exhibit high

degree of variability (Tsai *et al.* 1992). GSI systems S-loci have RNase activity, but not in *Papaver rhoeas* where small glycoproteins without RNase activity have been described (Franklin-Tong and Franklin 1993).

1.4.2. Structure and function of the pistil S proteins

The S-proteins in Solanaceae, Scrophulariaceae and Rosaceae are glycoproteins with one to nine N-linked glycan chains attached to each site of the protein. These N-linked carbohydrates may be involved in stability, conformation, and/or biological activity of the molecule (Ebert *et al.* 1989; Broothaerts *et al.* 1991; Oxley *et al.* 1998). Since these glycan chains have been found to be heterogeneous in number, type and structure, they could play a potential role in determining S-specificity (Woodward *et al.* 1989). Furthermore in animals the glycan chains of many glycoproteins have been shown to play an important role in cell-cell recognition (Lis and Sharon 1993).

To discover the role of the carbohydrate moiety in the S proteins, the effect of two different lectins on germination was tested. Lectins bind to saccharides, making them ineffective (Sharma and Shivanna 1983; Sharma *et al.* 1985). In these experiments inhibition of pollen germination was observed. These results were taken as evidence that the recognition molecules in the pistil contained saccharides and that binding of these with lectins made them ineffective in recognising self-pollen. However it should be noted that lectins inhibit any glycoprotein non-specifically. Therefore it is possible that the lectins used in these experiments inhibit other components of this system but not the S-proteins, which would give the same result (Karunanandaa *et al.* 1994). It is also known that the ribonuclease activity of the S protein is not affected by enzymatic

deglycosylation, so that the carbohydrate moiety may have other roles (Broothaerts *et al.* 1991). Direct evidence disproving the importance of these regions in determining S-specificity comes from transgenic experiments where S1S2 plants of *P. inflata* were transformed with a non-glycosylated S3 gene (in which the codon for the single N-glycosylation site asparagine was replaced with a codon for aspartic acid). These transgenic plants produce, in addition to the S1 and S2 proteins, the non-glycosylated S3 protein at wild-type levels. These plants reject not only S1 and S2 pollen but also S3 pollen (Karunanandaa *et al.* 1994). These results have been interpreted as evidence that the recognition function of these proteins resides in the amino acid backbone and not in the glycan chains. However only one experiment has been performed involving a single S allele with one glycan chain. Therefore, although these experiments are difficult to perform, it would be desirable to extend them to other alleles, so as to have further evidence concerning the role of the glycan chains in the determination of S-specificity, using different S-proteins with different numbers of glycan chains.

Sequences of the amino acid backbone of S-RNase alleles were first obtained for different solanaceous species. These revealed the presence of five short stretches of highly conserved regions, named C1 through C5, and two hypervariable regions, named HVa and HVb (Fig. 1.3; Ioerger *et al.* 1990; Tsai *et al.* 1992). The HVa and HVb regions are the most hydrophilic regions of the S-proteins (Ioerger *et al.* 1991). This led to the supposition that they may be on the external surface of the protein. The crystal structure of S-RNases has not yet been determined, and may vary for different S-RNases. However, based on the crystal structure of a fungus RNase similar to S-RNases, a model has been proposed where the two HV regions form a continuous surface on one

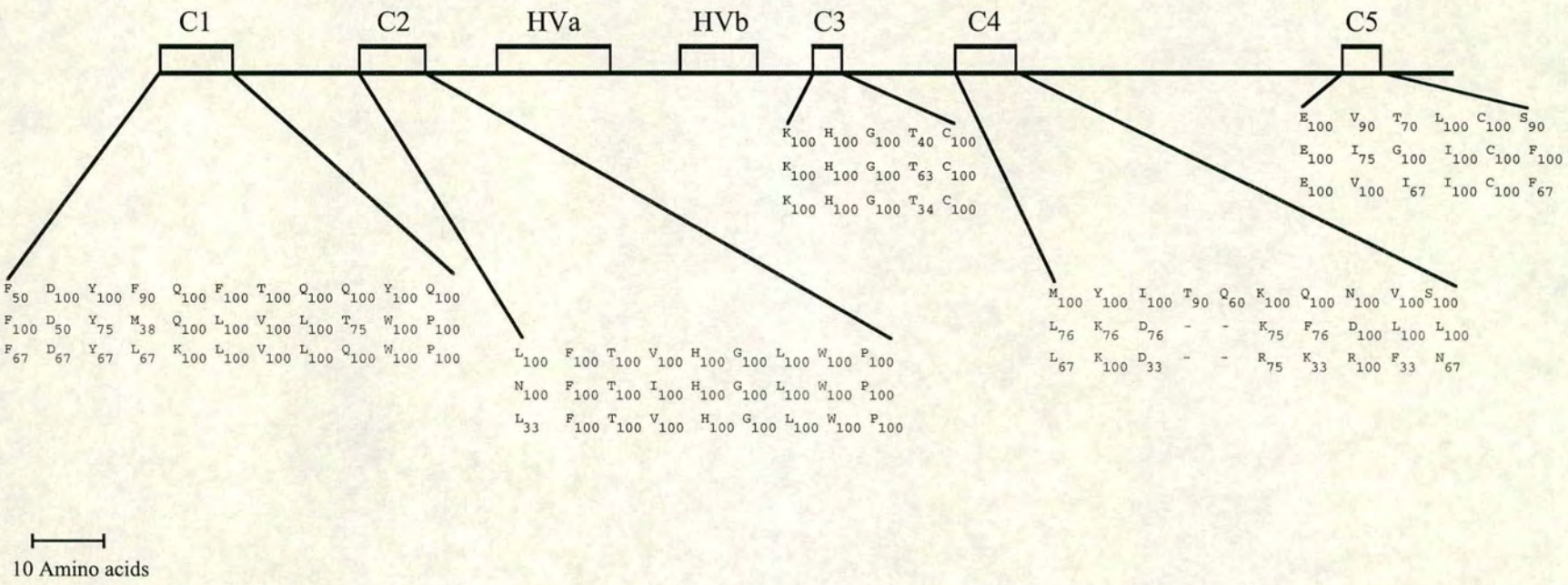


Fig. 1.3. Schematic representation of the structure of the S-locus. For the conserved regions consensus sequences (calculated from the data summarised by Richman *et al.* 1997) are presented for Rosaceae (top), Solanaceae (middle) and Scrophulariaceae (bottom).

site of the S-RNases (Parry *et al.* 1998). Therefore given their high degree of variability, these regions are candidates for being the ones determining S allele specificity. Also located in the HVa region is the only intron that varies in size (87 bp to 120 bp; Ebert *et al.* 1989; Jahnen *et al.* 1989; Kaufmann *et al.* 1991; Tsai *et al.* 1992; Coleman and Kao 1992; Saba-El-Leil *et al.* 1994).

What is (are) the region(s) that control(s) allelic specificity is still a matter of debate. The identification of the region(s) of the S-locus that determine(s) specificity has been difficult because of the extreme level of variation found both within and between species. Interspecific similarities are often as high as intraspecific similarities (Ioeger *et al.* 1990). This suggests that the polymorphism may be very ancient, consistent with the view that different alleles are maintained by balancing selection due to the frequency-dependent advantage of rare alleles. In only two instances have close enough alleles been found to allow us to draw some conclusions. In *Petunia,* two S-proteins with the same incompatibility type were found that differ at 13 amino acid positions (S2 and PS2; Clark *et al.* 1990). However, in *Solanum chacoense*, the nucleotide sequences of two alleles with different specificities (S11/ S13) have only ten amino acid differences, four of which are found in the HV regions, three of which are concentrated in HVa (Saba-El-Leil *et al.* 1994). Furthermore, sequence comparisons among S-RNases in Solanaceae revealed that nine scattered amino acids are highly variable (only one in the HV region; Fig. 1.3) and they could potentially play a role in pollen recognition (Tsai *et al.* 1992). The hypervariable regions have been proposed to be a component of the domain which confers allelic specificity to S proteins (Ioerger *et al.* 1991; Saba-El-Leil *et al.* 1994; Matton *et al.* 1997). The highest degree of sequence variation being in the HV regions has

been used as evidence to support the hypothesis that allelic specificity is determined by these regions. However it is possible that these regions are tolerant to structural variation and therefore variable without being necessarily involved in specificity determination (Zurek *et al.* 1997).

The conserved regions contain mostly hydrophobic amino acids and they are thought to be involved in the core structure of the S protein. The C2 and C3 regions are also found in fungal RNase enzymes and contain two of the three histidine residues that are required for catalytic activity of RNases of *Rhizopus niveus* (Kawata *et al.* 1990). In *P. inflata*, the replacement of the conserved histidine in the C3 region by an asparagine abolishes RNase activity. Pistils expressing this mutant S-allele fail to reject the corresponding S pollen (Huang *et al.* 1994), showing that RNase activity is necessary for the arrest of pollen tube growth.

Sequences of S-RNases in the families Rosaceae and Scrophulariaceae revealed that pistil S-RNases in these families have similar structures to solanaceous S-RNases (Fig. 1.3). No sequence related to the solanaceous C4 was identified in these families (Norioka *et al.* 1996; Sassa *et al.* 1996; Ushijima *et al.* 1998; Xue *et al.* 1996). S-RNases of Rosaceae with only one hypervariable region can be identified, corresponding to the solanaceous HVa (Norrioka *et al.* 1996; Sassa *et al.* 1996; Ushijima *et al.* 1998). As in solanaceous plants, the S-loci of most Rosaceae have one intron with different sizes (138 bp to 1100 bp; Broothaerts *et al.* 1995) in this region. However in sweet cherry (*Prunus avium*) at least two introns are present in four S-alleles (Tao *et al.* 1999).

This gene structure is not restricted to the S-RNase locus, other RNases called S-like RNases have the same structure (Ai *et al.* 1992; Jost *et al.* 1991; Löffler *et al.* 1992;

Löffler *et al.* 1993; Lee *et al.* 1992; Kuroda *et al.* 1994; Norioka *et al.* 1996; Ma and Oliveira 2000). Therefore S-RNases cannot be identified from their gene structure alone. Although S- and S-like RNases have similarities in structure they are distinct subclasses within the RNase superfamily (Green 1994). Unlike the S-RNases, S-like RNases are not flower-specific and have lower polymorphism levels than S-RNases. Since most of the S-like RNases have one intron located in the HVa region (Lee *et al.* 1992, Ma and Oliveira 2000) gene organisation cannot be used to distinguish both types of RNases. Therefore in order to identify S-alleles information on the gene structure is needed in combination with co-segregation and/ or polymorphism analyses.

1.5. Theoretical population models of the GSI

1.5.1 S-allele abundance in natural populations

It has long been recognised that both the pollen and the pistil components appear to be transmitted as a single Mendelian gene (East and Mangelsdorf 1925). This made possible the use of classical genetic crosses to estimate the number of distinct S-alleles in natural populations. It has been shown for several species that the number of S-alleles is very large, even in small populations (see as examples Emerson 1938; Lewis 1948; Lawrence 1975). Why are there so many S-alleles in natural populations? The answer lies in the consideration of the forces of mutation, selection and drift (see Clark 1993; Vekemans and Slatkin 1994). Suppose that initially three alleles (S1, S2, S3) are segregating in a population (these alleles have no other effects on fitness). Since these plants are SI, only heterozygotes are possible. If one allele becomes rare (say S1) then the population will consist mainly of genotype S2S3. Therefore there is an advantage for

the S1 pollen over the other two, because it can fertilise most plants in the population, and it will thus increase in frequency. However as its frequency increases its advantage over the other two alleles disappears. Therefore in this system, allelic diversity is not only advantageous, it is necessary. At equilibrium all three alleles will have equal frequency in the population. Similarly when a mutation that generates a new allele with a new specificity (S4) occurs, its frequency will rapidly increase until it reaches a frequency equal to that of the other alleles. In a finite population a balance between mutation-selection sweeping in new alleles and loss of alleles by drift maintains this equilibrium. This type of selection is referred to as frequency dependent selection (reviewed by Clark 1993).

1.5.2. Ancient polymorphism of the S-alleles

Because S-alleles are maintained in the population by frequency dependent selection they persist for longer times than expected under neutrality. Therefore these alleles are expected to be old and very distinct even at silent and intron sites. The availability of molecular sequence data from distantly related Solanaceae species (for instance the most recent common ancestor of *Nicotiana* and *Petunia* is thought to have lived approximately 70 million years ago (MYA); Wolf *et al.* 1989) makes it possible to test this prediction. The S-allele genealogies in Solanaceae have revealed that alleles from different species can cluster in the same branch since some interspecific similarities are higher than intraspecific similarities, suggesting that some polymorphisms are more ancient than species divergences (Ioerger *et al.* 1990). However these shared polymorphic sites could be the result of either common ancestry or they could have

arisen independently in each lineage after species divergence. If these shared polymorphic sites are of common ancestry, an excess number of nucleotide sites that are segregating for the same pair of nucleotides over the number of segregating sites that are created by chance is expected (Ioerger *et al.* 1990). Computer simulations were used to obtain the number of shared polymorphisms obtained by chance that is compatible with the observed polymorphism level in the S-alleles. The numbers of shared polymorphisms obtained in the computer simulations were always less than the number observed. Therefore the excess of shared polymorphisms is consistent with these being the result of common ancestry (Ioerger *et al.* 1990).

In order to obtain sequences as divergent as S-alleles, there must either be a very high mutation rate, or else a very long history of independent divergence. Population genetics theory predicts that the genealogy of alleles under frequency dependent selection has the same structure as alleles at a neutral locus, but the time scale of the genealogy under selection is expanded (Clark 1993; Vekemans and Slatkin 1994). Monte Carlo simulations show that at higher mutation rates a neutral gene gains more alleles, but the expected time to coalescence of all the alleles remains $4N_e$ generations (Clark 1993; Vekemans and Slatkin 1994). The S-allele genealogy decreases in coalescence time as the mutation rate increases. Therefore as the mutation rate increases, the S-allele genealogy looks more like a neutral genealogy (Clark 1993; Vekemans and Slatkin 1994; Clark 1996).

1.5.3. Detecting selection on S-alleles

The first two population studies of the GSI system were on one population of *Solanum carolinense* (Richman *et al.* 1995) and another of *Physalis crassifolia* (Richman *et al.* 1996a). These provided the first extensive evidence on the level of DNA sequence polymorphism of S alleles in natural populations. However the incompatibility types of the alleles sequenced are not known (Richman *et al.* 1996a; 1996b). These studies confirmed earlier findings that alleles differ at multiple sites, and they yield the best estimates of the average number of substitutions between S alleles. However the two species differ in both allele numbers and apparent ages of the alleles. The degree of sequence variation in *S. carolinense* suggests that these allelic lineages are extremely ancient. In *P. crassifolia,* the S-allele sequences are more similar to one another, suggesting a less ancient origin, perhaps due to a relatively recent population bottleneck (Richman *et al.* 1996b). Further population studies in three other species of Solanaceae (*P. cinerascens*, *Lycium andersonii* and *Witheringia maculata*) have also confirmed that the number and age of alleles varies among species (Richman and Kohn 1999, 2000). These authors suggested that these differences are the result of ecological differences among these species. A detailed understanding of how S-alleles are maintained in natural populations may therefore require data on demographic and ecological parameters as well.

The region that includes the HV regions of the Solanaceae S-alleles was analysed in detail (Richman *et al.* 1996b). Alleles that are similar in sequence tend to have higher ratios of non-synonymous to synonymous changes than do more distant alleles. The average ratio of non-synonymous to synonymous changes is slightly biased towards

synonymous differences (Richman *et al.* 1996b; reviewed by Charlesworth and Guttman 1997). This pattern has also been found in other loci involved in recognition processes (Lee and Vaquier 1992; Hughes *et al.* 1990). The pattern of an excess of amino acid substitutions among closely related alleles suggests balancing selection, as follows (Hughes and Nei 1989). Once a new specificity type is established, the presence of conserved sites, and sometimes restrictions in the amino acids allowed at variable sites, probably limits further protein sequence divergence (Tanaka and Nei 1989), and this could explain the average slight bias towards synonymous differences. The average number of nonsynonymous differences between allele pairs increases more slowly than synonymous differences (Uyenoyama 1997). This is the opposite of what occurs with saturation at synonymous sites and suggests diversifying selection, as differences accumulating neutrally should show a linear relation (Tanaka and Nei 1989). This effect might help test which regions are under diversifying selection. However the capacity to detect functionally different regions in genes where variant alleles have persisted for very long time periods, as with the S-alleles depends on the occurrence of recombination (see Awadalla and Charlesworth 1999).

1.6. Chapter Outlines

In **Chapter 2**, I present the rationale for choosing the *Antirrhinum* and *Misopates* genera to study the effect of the breeding system on levels of variation, and to study the S locus.

In **Chapter 3**, I characterise the mating system of nine *Antirrhinum* and three *Misopates* populations, since it is known that the breeding system is a characteristic of a population not of the species as a whole. In addition, four morphometric characteristics have been measured in order to determine their utility as selfing-rate indicators.

To study the effect of the breeding system on levels of genetic diversity I studied the cycloidea and *fil1* genes (**Chapter 4** and **5**). These genes, however, have unexpectedly little diversity within species and divergence between species and therefore are unsuitable for the proposed work. Much of the work presented in **Chapter 4** and **5** has been published in Molecular Biology and Evolution by Vieira *et al.* (1999) and Journal of Molecular Biology by Vieira and Charlesworth (2000). In **Chapter 6**, I determine whether the low diversity and divergence is a general pattern, by determining levels of divergence between *Antirrhinum majus* and *Verbascum nigrum* for 5 additional genes (*fil2-1*; *fil2-2*; *farS*; *globosa1*; *Adh*). The work presented in this chapter has been submitted to Plant Molecular Biology by Vieira and Charlesworth.

In **Chapter 7**, I characterise at the molecular level variability in the hypervariable regions of several S-alleles of *Antirrhinum* and *Misopates,* the regions that have been proposed to be involved in specificity determination.

Chapter 8 summarises the main findings in the thesis and elaborates on further research directions.

Chapter 2

Choice of the genus

The snapdragon, *Antirrhinum majus* (Scrophulariaceae), is one of the classical species for genetic studies. The short vegetative period, the small size of the plants, the fact that they are easy to cultivate, the fast germination (7 to 10 days), and the size and flower structure are some of the characteristics that make this organism suitable for genetic research (Harte 1974). Species of this genus were first studied by Godron in 1863, long before the rediscovery of Mendel's laws. However its history as a research organism for genetics really begins with the studies of Wheldale (1907) on the inheritance of the flower colour variation and the studies of Baur (1930), who characterised many mutants.

Numerous studies on transposable elements (Bonas *et al.* 1984; Sommer *et al.* 1985; Upadhyaya *et al.* 1985; Coen and Carpenter 1986) led to the discovery of several highly mobile transposable elements. Among other uses, these were useful to develop a system of transposon-mutagenesis (Martin *et al.* 1985; Sommer and Saedler 1986; Martin *et al.* 1991) that can be used to clone and sequence genes responsible for a

particular phenotype, as was done for several genes involved in flower development (Carpenter and Coen 1990, Schwarz-Sommer *et al.* 1990, 1992). Currently, 75 genes have been partially or completely sequenced in this model organism (GenBank release version 114.0, 08/09/00). The majority of these genes are homeotic genes or genes involved in floral pathways.

The genus *Antirrhinum* has a Mediterranean range, most of the species living in the Iberian Peninsula, which is the diversification centre of the genus (Caputo *et al.* 1991). Most of the species are endemic plants that live in fragmented areas with small population sizes (Anon. 1983, 1992; Barreno *et al.* 1984; Laguna 1994).

The taxonomy of the genus based on morphological characters is still unresolved (Doaigey and Harkiss 1991). Several species have changed status according to different authors. The first description of the genus *Antirrhinum* by Linnaeus (1753, in Sutton 1988), included species now classified as *Antirrhinum*, *Linaria* and *Elatine*. Several authors reclassified *Antirrhinum* into several sections with different numbers of species. Rothmaler (1956) distinguished *Antirrhinum* based on the absence of a corolla spur (present in *Linaria* species) and the length of the calyx (longer than the corolla tube in *Misopates* species). Some *Antirrhinum* subspecies are still regarded by some authors as different species, and vice-versa. Therefore, the number of species that are included in this genus is constantly changing. We follow the classification of Franco (1971) for the Portuguese species, and for the remaining European species the classification of Halliday and Beadle (1983). According to these authors, the genera *Antirrhinum* and *Misopates* are composed, respectively, of twenty two perennial species and three annual species (Mather 1947; Doaigey and Harkiss 1991; Güemes *et al.* 1993).

The close juxtaposition of the anthers and stigma in the flowers of these genera means that self-pollination is likely to occur which would result in self-fertilisation in the absence of a preventive mechanism. A self-incompatible gametophytic system seems to be present in many *Antirrhinum* species, except for *A. siculum*, cultivated varieties of *A. majus* and all *Misopates* species (Baur 1911, 1919; Lotsy 1911; Sherman 1939; Harrison and Darby 1955; Xue *et al.* 1996). However, it is known from other taxa that individuals of different populations of the same species may sometimes have different breeding systems (e. g. *Leavenworthia crassa*, Lloyd 1965; *Lycopersicon pimpinellifolium*, Rick *et al.* 1977; *Lycopersicon hirsutum*, Rick *et al.* 1979; *Gilia achilleifolia*, Schoen 1982; *Eichhornia paniculata*, Glover and Barret 1986; *Turnera ulmifolia* Barrett and Shore 1987; *Clarkia tembloriensis*, Holtsford and Ellstrand 1989; *Mimulus guttatus*, Ritland and Ritland 1989).

All species of these genera that have been studied are diploid (2N=16) and have 2 large, 5 medium and 1 small chromosomes that can be distinguished from one another in the pachytene phase (Harte 1974; Fernandes *et al.* 1977).

In experimental crosses, the *Antirrhinum* species easily hybridise, giving viable progenies (Mather, 1947; Rothmaler, 1956). This seems to be a common event among Scrophulariaceae (Elisens 1992). Among *Antirrhinum*, only *A. siculum* has been described as genetically isolated (Rothmaler 1956). However, few natural hybrids have been described (Sutton 1988), suggesting that hybridisation in nature is rare. This may be due to different geographic distributions (different species rarely share the same habitat), and different pollinators (Mather 1947, Kampny 1995).

It is often useful to compare levels of within species variability to species divergence. In my work I have used as outgroups species of the *Digitalis* and *Verbascum* genera. For most of the Scrophulariaceae phylogenetic relationships at the genus level are not known, because plant molecular evolution has been dominated by studies of chloroplast genes. Because of their slow rate of evolution they are most suitable for studies of plant phylogenetic relationships at or beyond the family level (Wolfe *et al.* 1989; Clegg *et al.* 1997). Only one study using 18S ribosomal nuclear genes has been performed in the Scrophulariacea family (Soltis *et al.* 1999). In this study there is no power to distinguish the relationships of the Scrophulariaceae genera. Nevertheless, the chloroplast sequence data supports the hypothesis that *Antirrhinum* and *Digitalis* belong to one Scrophulariaceae clade, while *Verbascum* belongs to a different one (Olmstead and Reeves 1995; Wolfe and dePamphilis 1998; Soltis and Soltis 2000). However bootstrap values supporting these two clades are always below 74%. When any of the three chloroplast genes (*ndh*F, *trn*L, and *rbc*L) is used, a higher number of synonymous differences per synonymous site is obtained in the comparison involving *Antirrhinum* and *Digitalis* than in the *Antirrhinum* and *Verbascum* comparison (Table 2.1). Therefore it is still unclear whether *Antirrhinum* is more closely related to *Digitalis* than to *Verbascum*.

Table 2.1. Silent site divergence for three chloroplast genes between *Antirrhinum* and both *Digitalis* and *Verbascum*.

Gene	Region analysed (bp)	*Antirrhinum* compared with	Silent site divergence
*ndh*F	2093	*D. purpurea*	0.2506
		V. thapsus	0.1751
*trn*L	518	*D. purpurea*	0.0724
		V. thapsus	0.0531
*rbc*L	1329	*D. purpurea*	0.1339
		V. thapsus	0.0865

Accession numbers for *ndh*F are: *A. majus* L36392, *D. purpurea* AF130150, *V. thapsus* L36417; for *trn*L are: *A. majus* AF118790, *D. purpurea* AF034871, *V. thapsus* AF118804; and for *rbc*L are: *A. majus* L11688, *D. purpurea* X83720, *V. thapsus* L36452.

The large number of closely related species in the *Antirrhinum* and *Misopates* genera, their differences in the mating system, coupled with *Antirrhinum majus* being a model species, makes this taxonomic group a logical choice for molecular studies addressing the effect of the breeding system on overall genetic diversity and in paticular on GSI.

Chapter 3

The breeding system and variation in four morphometric characters of *Antirrhinum* and *Misopates* populations

3.1 Introduction

Evolutionary forces such as mutation, selection, migration, genetic drift, recombination, effective population size and breeding system shape patterns and levels of genetic diversity (Aquadro and Begun 1993; Tajima 1993; Moriyama and Powell 1996; Charlesworth *et al.* 1997; Clegg 1997; Clegg *et al.* 1997; Amos and Harwood 1998; see Chapter 1). In plants, mating systems range from regular systems of inbreeding by self fertilisation, breeding among relatives within small neighbourhoods, to strict outcrossing maintained by incompatibility systems (Stebbins 1957, Barrett and Eckert 1990). In completely selfing populations, only loci with symmetrical selection coefficients against the homozygous genotypes will remain variable, and such loci are presumably rare (Kimura and Ohta 1971). Ultimately, inbreeding increases homozygosity and therefore the effectiveness of selection against recessive deleterious alleles, resulting in a purge of detrimental alleles from the population and a higher mean

fitness of the population. However, this process takes time and the initial effect of an increase in inbreeding is a reduced mean fitness of the population due to homozygosity (inbreeding depression; Charlesworth and Charlesworth 1987; Barrett and Charlesworth 1991).

The most common empirical descriptor of the mating structure of plant populations is the outcrossing rate, the proportion of offspring fathered by individuals other than their seed parent. The outcrossing rate of a given genotype results from the interaction of ecological factors, e.g. pollination abundance and plant density (Schemske and Lande 1985); with genetically determined traits. Characters such as flower size, flower morphology, protandry (i.e, the amount of time separating anther dehiscence and the start of stigma exsertion), anther-stigma separation, and the ability to set seeds in the absence of pollinators are examples of genetic factors that affect outcrossing rates.

Depending on the species, different floral features are indicators of the breeding system. Reduction in flower size is usually, but not invariably, associated with self-pollination in flowering plants (Moore and Lewis 1965). For instance, in the genus *Clarkia* outcrossers have large flowers and long styles, and anthers are bigger than in selfers. However the self-pollinating species *Clarkia tenella* has flowers as large as those of outcrossing species (Moore and Lewis 1965). Only in outcrossing species is it advantageous to allocate resources to conspicuous flowers. Therefore in highly selfing populations small flowers may represent the result of relaxed selection for conspicuous flowers by pollen vectors coupled with positive selection for small flowers and flower parts (excluding the ovary; Moore and Lewis 1965). Protandry is also an important determinant of the breeding system in many plants, for example *Nicotiana rustica*

(Breese 1959), *Clarkia* (Moore and Lewis 1965), *Limanthes* (Arroyo 1973), and *Gilia achilleifolia* (Schoen 1982). Anther length and stigma exsertion are positively correlated with outcrossing rates in *Lycopersicon pimpinellifolium* (Rick *et al.* 1977). Stigma-anther separation is positively correlated with outcrossing rates in *Mimulus guttatus* (Ritland and Ritland 1989) but in *Gilia* no significant association was found (Schoen 1982).

Environmental variation can also affect outcrossing rates. Both ecological and demographic factors influence the process of mating. The influence of ecological factors is observed, for instance, in *Collinsia sparsiflora* var. *arvensis*, in which outcrossing depends strongly on flower density (Kahler, personal communication, in Schemske and Lande 1985). In *Lupinus succulentus* and *L. nanus* variation in mating system is mostly due to fluctuations in bee activity (Schemske and Lande 1985).

The reliability of the floral features as indicators of the breeding system may depend on population size. For instance in *Clarkia tembloriensis* long styles are part of a large flowered outcrossing syndrome, and therefore style length has been described as a convenient index of outcrossing. However this appears to be a less reliable indicator when small populations are scored (Vasek and Hording 1976). Therefore inferences about plant breeding systems based solely on floral morphological grounds, although informative, may sometimes be misleading, and it is desirable to also determine the mating system experimentally.

I established the mating system in the glasshouse of the populations and species studied here by estimating the percentage of autogamy and self-fertility. Two species of *Misopates* and five of *Antirrhinum* (including multiple populations of several species;

Table 3.1) were analysed. Since floral and mating system evolution are likely to be correlated for these populations, four floral morphometric characters were also measured.

3.2. Material and Methods

Establishing the mating system

The mating system was estimated in the glasshouse for the populations presented in Table 3.1. Ten seeds were sown from at least three fruits from a number of different wild plants. For every plant grown, five flowers were allowed to self-pollinate without manipulation, to estimate the percentage of autogamy (fruit set by selfing by within-flower pollination; Richards 1997). All flowers were bagged to prevent receipt of pollen from neighbouring plants. For the same plants, five flowers were hand-pollinated with their own pollen, to estimate the production of seeds as a measure of self-fertility. In order to rule out apomixis (asexual reproduction; Richards 1997) for the apparently self-compatible plants, I removed the anthers in ten flowers of each plant three days before anthesis.

Floral variation in the *Antirrhinum* populations

For each of the populations four floral features were measured on five flowers of every plant grown. Illustrations of flowers of each taxa are presented in Figure 3.1 and 3.2. The *Antirrhinum* corolla is gamopetalous, tubular, and bilabiate. The abaxial lip has a basal convexity termed the "palate" which occludes the mouth of the tube (Fig. 3.2, Sutton 1988). The floral features measured were: corolla length (from the base of the corolla tube to the abaxial lip), corolla diameter (in the region of the prominent basal palate of the

Table 3.1. Species, population codes, collection localities, and years of collection.

Species	Population code	Locality	Collection year
Antirrhinum			
majus cirrhigerum	cirrhigerumAve	Portugal- Aveiro, Praia da Barra	1997
	cirrhigerumGala	Portugal- Figueira da Foz- Gala, Praia do Hospital	1998
	cirrhigerumMuel	Portugal- Marinha Grande, Sao Pedro de Muel	1998
	cirrhigerumAlg	Portugal- Algarve	
majus majus	majus	France- Brissac and St-Etienne d'Issensac	1998
majus linkianum	linkianum	Portugal- Coimbra, road Pombal-Ansiao	1998
latifolium	latifolium	France-Usson-Les-Bains	1998
molle lopesianum	molle	Portugal- Braganca, Alfeiao	1998
meonanthum	meonanthum	Portugal- Vila Real, Regua	1998
graniticum	graniticumB	Portugal- Braganca, Monte de Sao Bartolomeu	1997
Misopates orontium	orontiumB	Portugal- Braganca, Monte de Sao Bartolomeu	1997
	orontiumG	Portugal- Vila Nova de Gaia, Praia do Areinho	1997
calycinum	calycinum	Portugal- Coimbra	1998

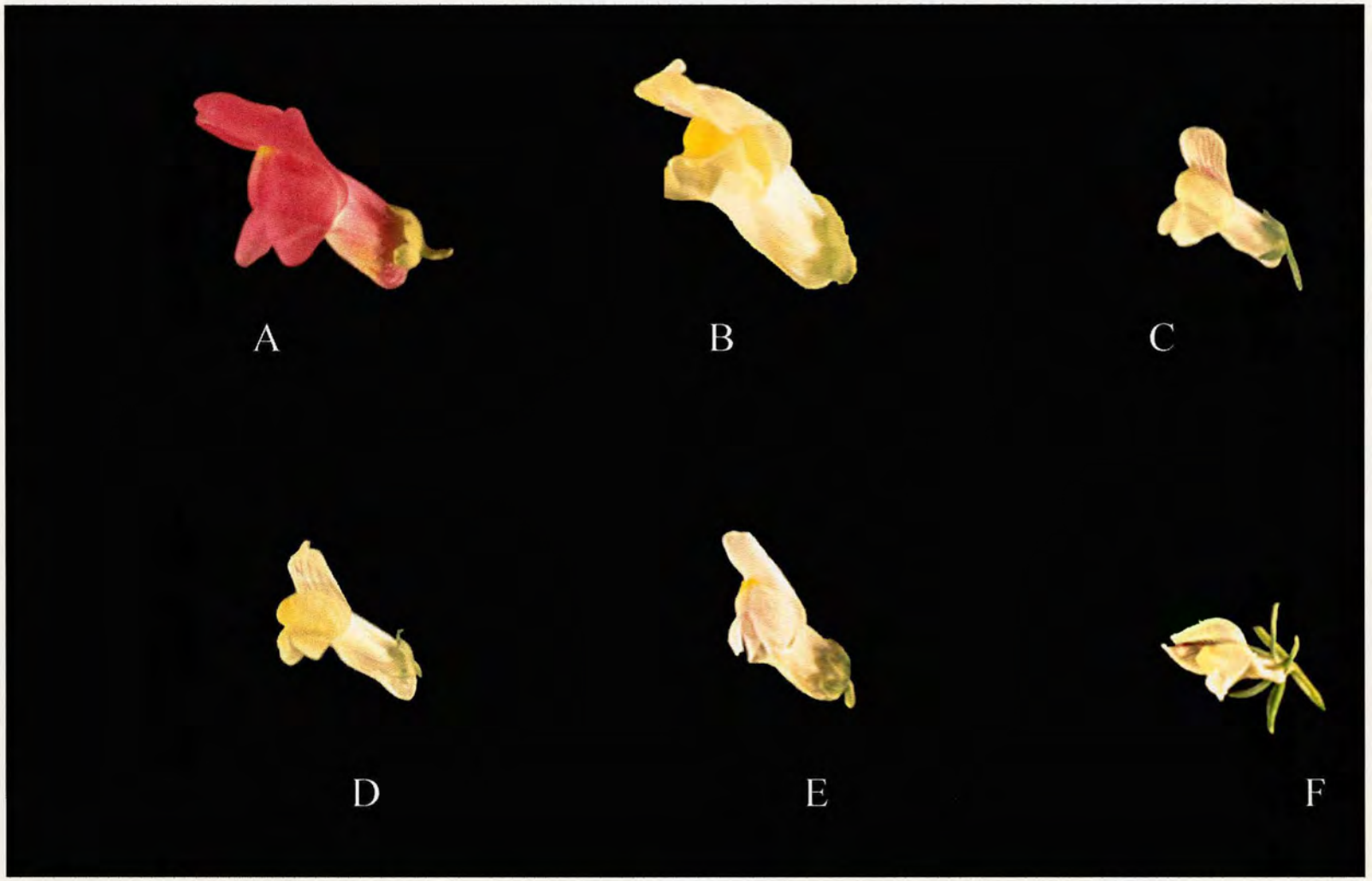


Fig. 3.1. Flowers of five *Antirrhinum* and one *Misopates* species: A- *A.majus* subsp. *cirrhigerum*, B- *A. latifolium*, C- *A. molle*, D- *A. meonanthum*, E- *A. graniticum*, F- *Misopates calycinum.*

abaxial lip), and the distance between both the short and long anther and the stigma two days after anthesis (Fig. 3.2). The floral traits were measured in a common environment, therefore the differences observed should be predominantly genetically based (unless there are large nongenetic maternal effects on these characters).

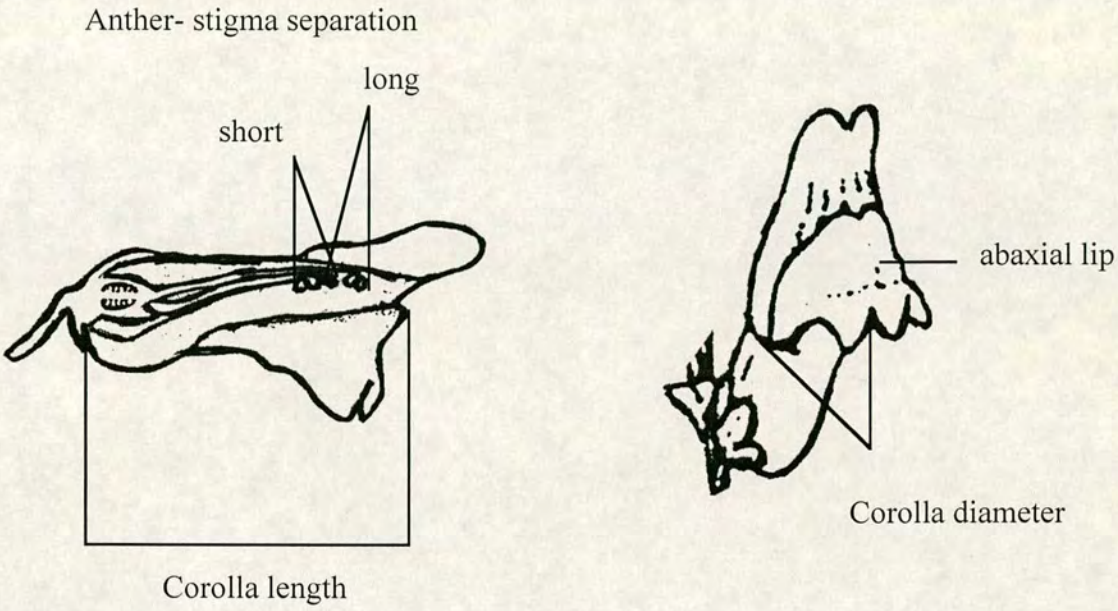


Fig. 3.2. Measurements made on flowers of *Antirrninum* and *Misopates.*

3.3. Results

3.3.1. The mating system

The mating system results are presented in Table 3.2. Some individuals did not produce seeds after hand self-fertilisation. For these I ruled out the possibility of non-functional pistils and/ or pollen (i. e. female and male sterility), since they produce fruit set when crossed with other individuals (Table 3.3). After removal of the anthers prior to anthesis none of the self-compatible individuals analysed set seeds and thus apomixis is ruled out (Table 3.4).

Table 3.2. Results of the autogamy and self-fertility tests on plants of the *Antirrhinum* and *Misopates* populations.

Species	Population code	Number of maternal plants	Number of plants analysed	Autogamy (%)	Self-fertility (%)	Mating system
Antirrhinum						
A. majus subsp *cirrhigerum*	cirrhigerumAve	9	35	96	95.4	largely self-compatible
	cirrhigerumGala	4	13	80.0	78.4	largely self-compatible
	cirrhigerumMuel	4	17	7.06	4.71	largely self-incompatible
	cirrhigerumAlg	2	2	0	0	Self-incompatible
majus	majus	2	6	0	0	Self-incompatible
linkianum	linkianum	6	11	25	28.3	partially self-incompatible
A. latifolium	latifolium	2	7	5.71	8.57	largely self-incompatible
A. molle	molle	4	9	0	0	Self-incompatible
A. meonanthum	meonanthum	3	9	0	0	Self-incompatible
A. graniticum	graniticum	2	5	0	0	Self-incompatible
Misopates						
M. orontium	orontiumB	4	12	100	100	Self-compatible
	orontiumG	3	8	100	100	Self-compatible
M. calycinum	calycinum	3	25	100	100	Self-compatible

Table 3.3. Male and female fertility in individuals that did not produce seeds after hand self-fertilisation. Number of cross-pollinated flowers which produced seeds / total number of flowers pollinated. Pollinations were performed with several individuals of the same population. Dash indicates pollination not performed.

Population	Plant code	Pollinations pollen donor	pistil receptor	Population	Plant code	Pollinations pollen donor	pistil receptor
cirrhigerumAve	5.2	2/2	2/2	linkianum	3.1‡	3/8	4/7
cirrhigerumGala	3.2	4/4	3/3		12.1	7/7	3/5
	4.3	3/3	3/3		12.2‡	2/5	3/8
cirrhigerumMuel	1.1	8/9	9/9	molle	3.2	2/3	4/8
	1.2	11/16	16/17		3.3	1/1	4/4
	1.3	15/18	15/15		7.1	2/2	3/3
	1.4	13/14	11/13		7.2	3/3	2/2
	2.1	24/26	15/18		10.1	5/7	-
	2.3	13/13	9/9		10.3	4/4	3/3
	4.1	23/23	21/24		16.1	7/7	6/6
	4.2	10/11	13/14	meonanthum	2.1	22/30	21/25
	4.3	19/20	15/18		2.2	35/39	23/28
	4.4	17/18	18/18		2.3	30/31	28/34
	4.5	8/8	15/15		3.1	36/41	23/32
cirrhigerumAlg	1	4/4	4/4		3.2	45/62	28/29
	2	4/4	4/4		3.3+	4/12	32/39
majus	1.1*	18/21	13/17		5.1+	22/29	16/36
	1.2*	12/17	11/14		5.2	7/7	29/29
	2.1	15/15	12/15	graniticum	1.1	11/11	8/12
	2.2	10/12	19/19		1.2	13/13	11/12
	2.3	20/21	14/14		4.1	3/6	11/11
	2.4	11/12	17/19		4.2	7/10	2/3
					4.3	5/5	5/7

Pollinations using individuals that share the same symbol (*, ‡, and +) are cross-incompatible.

Table 3.4. Summary of tests for apomixis in the individuals that produced seeds after within-flower pollination and hand self-fertilisation.

Species	Population	Number of plants studied	Number of cases that produced seeds
Antirrhinum majus			
	cirrhigerumAve	34	0
	cirrhigerumGala	11	0
	cirrhigerumMuel	2	0
Misopates			
	orontiumB	12	0
	orontiumG	8	0

3.3.2. Morphological variation

Table 3.5 and 3.6 show the flower character data. It can be seen that the coefficients of variation for corolla length (except *M. orontium*) and diameter are lower than for both measures of anther-stigma separation. Therefore the corolla characters seem to be less plastic than the anther-stigma separation.

Based on each of the four morphometric variables, the six *Antirrhinum majus* populations are significantly different (Table 3.7). The four populations of *Antirrhinum majus* subsp. *cirrhigerum* are also significantly different.

Table 3.5. Corolla length and diameter (cm) in *Antirrhinum* and *Misopates* populations

Population code	Number of flowers analysed	Corolla length	Coefficient of variation %	Corolla diameter	Coefficient of variation %
cirrhigerumAve	107	2.771 ± 0.226	8.156	0.986 ± 0.084	8.519
cirrhigerumGala	67	2.906 ± 0.160	5.506	1.026 ± 0.084	8.187
cirrhigerumMuel	89	2.748 ± 0.147	5.349	1.019 ± 0.061	5.986
cirrhigerumAlg	10	2.748 ± 0.142	5.167	0.919 ± 0.064	6.964
majus	31	3.249 ± 0.221	6.802	1.163 ± 0.109	9.372
linkianum	61	2.769 ± 0.140	5.056	0.945 ± 0.094	9.947
latifolium	34	3.218 ± 0.163	5.065	1.113 ± 0.074	6.647
molle	43	1.847 ± 0.170	9.204	0.681 ± 0.064	9.398
meonanthum	45	2.172 ± 0.196	9.024	0.602 ± 0.079	13.120
graniticumB	25	2.187 ± 0.112	5.121	0.681 ± 0.069	10.132
orontium	5	0.970 ± 0.478	49.278	0.378 ± 0.025	6.614
calycinum	134	1.585 ± 0.086	5.426	0.508 ± 0.038	7.480

Table 3.6. Short and long anther-stigma separation (cm) in *Antirrhinum* and *Misopates* populations

population code	Number of flowers analysed	Short anther-stigma separation	Coefficient of variation %	Long anther-stigma separation	Coefficient of variation %
cirrhigerumAve	107	0.290 ± 0.079	27.241	0.094 ± 0.079	84.043
cirrhigerumGala	67	0.231 ± 0.109	47.186	0.081 ± 0.130	160.494
cirrhigerumMuel	89	0.251 ± 0.086	34.263	0.056 ± 0.099	176.786
cirrhigerumAlg	10	0.318 ± 0.124	38.994	0.124 ± 0.056	45.161
majus	31	0.165 ± 0.132	80.00	0.290 ± 0.127	43.793
linkianum	61	0.150 ± 0.135	90.00	0.173 ± 0.130	75.145
latifolium	34	0.300 ± 0.142	47.333	0.074 ± 0.109	147.297
molle	43	0.224 ± 0.048	21.429	0.0	
meonanthum	45	0.150 ± 0.089	59.333	0.091 ± 0.091	100.0
graniticumB	25	0.221 ± 0.091	41.176	0.147 ± 0.091	61.905
orontiumB	5	0.0		-0.109 ± 0.038	-34.862
calycinum	134	0.150 ± 0.043	28.667	0.008 ± 0.041	512.5

Table 3.7. Ratio of the between and within population variance (F ratio) for four flower characters

Species	Number of populations	Corolla length	Corolla diameter	Anther-stigma separation short	Anther-stigma separation long
A. majus	6	43.201***	33.953 ***	19.065 ***	27.012 ***
A. majus subsp. *cirrhigerum*	4	10.700 ***	9.317 ***	7.828 ***	2.872 *

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$

Across populations and species, the morphometric variables were not independent (Fig. 3.3A - E). There are significant correlations among all of these except between the upper and the lower anther-stigma separation values. Across *A. majus* populations a significant correlation is only found between corolla lenght and diameter (Fig 3.4).

When all populations and species are included, or when just *A. majus* populations are considered, none of the morphometric variables here studied are significantly correlated with the breeding system (Table 3.8). Therefore, they cannot be used as indicators of the breeding system.

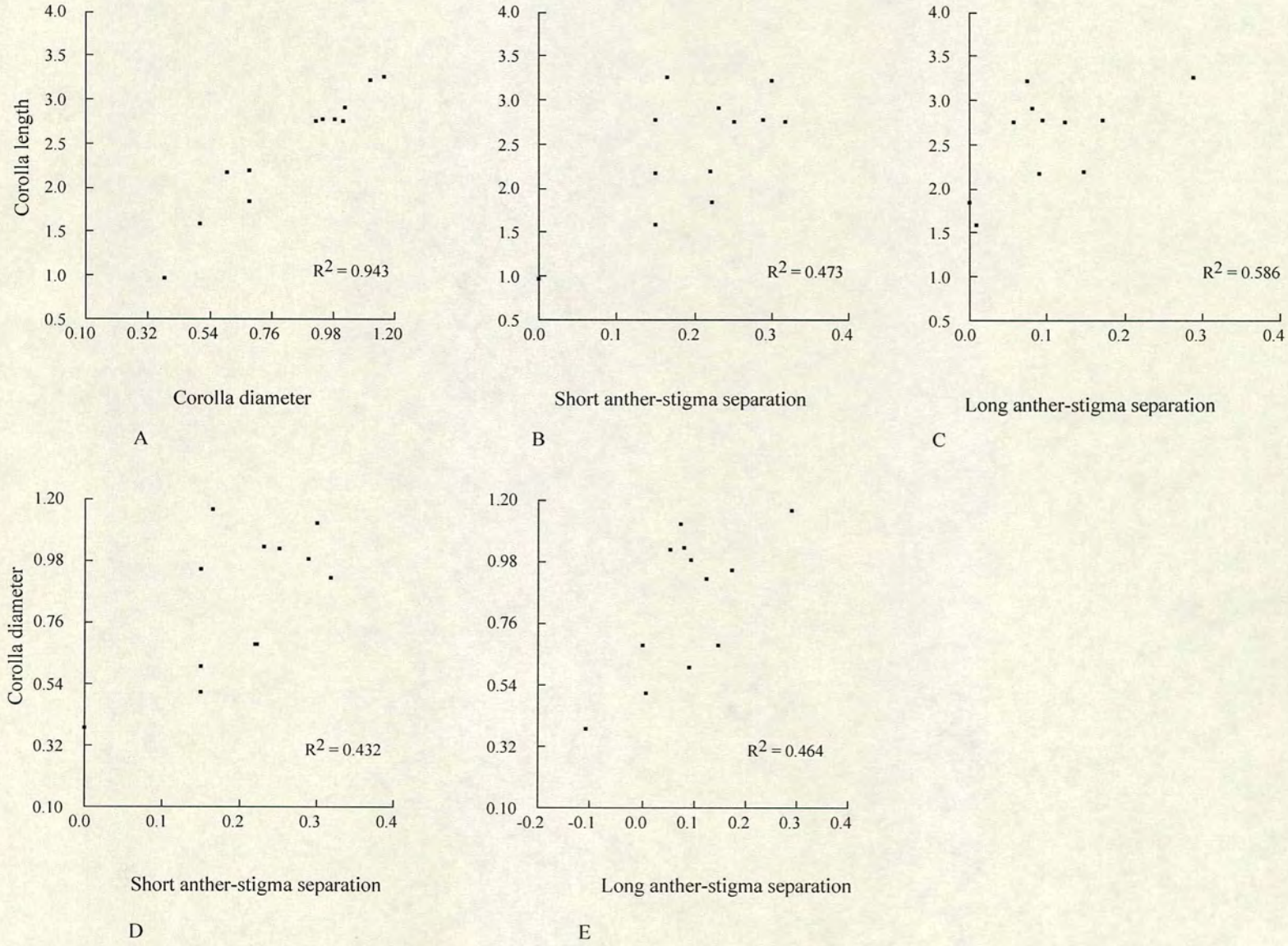


Fig. 3.3. Correlations of the average flower characters values in twelve *Antirrhinum* and *Misopates* populations.

Table 3.8. Variance explained by the linear regression of each of the four morphometric variables on self-fertility.

	Number of populations	Corolla		Anther-stigma separation	
		length	diameter	short	long
All populations and species	12	0.18	0.12	0.14	0.26
A. majus populations	6	0.05	0.02	0.05	0.20

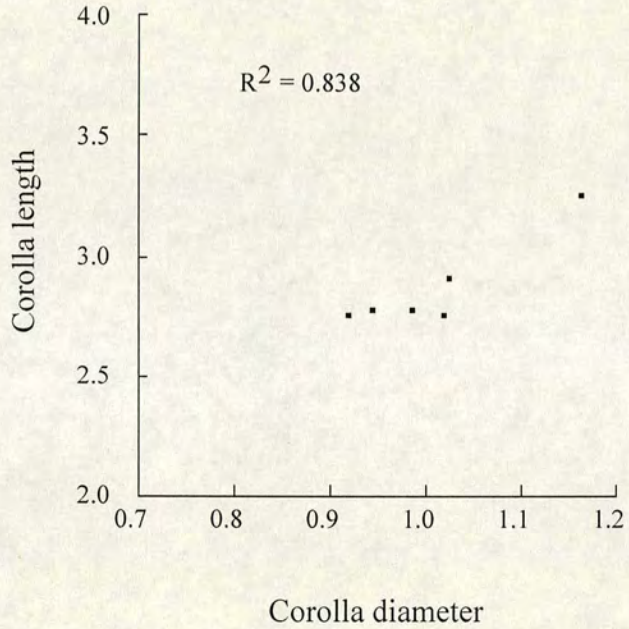


Fig. 3.4. Correlation of the average length and diameter of the corolla in six *Antirrhinum majus* populations

3.3.3. Within and between family variance

In a population of selfers, the within-family genetic variance should approach zero, but there should be a remainder attributed to environment. The genetic variance will then be attributable entirely to the between-family component. Thus for self-compatible

populations, the between-family component should be larger than in self-incompatible populations (Solbrig and Rollins 1977, Lawrence 1984). There is no clear pattern in our *Antirrhinum* and *Misopates* data (Table 3.9). The between family component of the variance is greater than the within family component of variance for many characters for both self-compatible and self-incompatible species.

Table 3.9. Ratio of the between and within families variance (F ratio) for four flower characters

Population code	Number of families	Corolla length	Corolla diameter	Anther-stigma separation short	Anther-stigma separation long
cirrhigerumAve	8	4.958 **	4.029 **	2.715 *	5.876 ***
cirrhigerum Gala	4	8.318 ***	13.837 ***	4.025 *	3.798 *
cirrhigerumMuel	4	33.19 ***	1.974	7.45 ***	3.493 *
majus	2	3.428	1.825	15.102 **	11.766 **
linkianum	6	1.311	23.083 ***	9.749 ***	7.681 ***
latifolium	2	0.739	0.209	9.015 **	8.201 **
molle	4	6.676 **	1.245	2.513	-
meonanthum	3	2.046	8.105 **	0.289	2.78
graniticumB	2	9.695 **	9.010 **	0.442	12.859 **
calycinum	3	0.127	0.090	11.738 ***	3.293 *

* $P < 0.05$; ** $P < 0.01$; *** $P < 0.001$

3.4. Discussion

In the genus *Antirrhinum*, most species are described as being self-incompatible, but *A. majus* and *A. siculum* are said to be self-compatible (Harrison and Darby 1955; Herrmann 1973; Sutton 1988; Xue *et al.* 1996). *Misopates* species have been described as self-compatible (Mahal and Pal 1984). For one species of this genus, *Misopates orontium,* flowers become smaller and cleistogamous late in the flowering season, and therefore the outcrossing rate must become small. As expected the populations of *Misopates* were found to be self-compatible. The mating system in the *Antirrhinum* species varies from largely self-compatible to self-incompatible. Within *A. majus,* different populations have different breeding systems. This is true even within the subspecies *A. majus cirrhigerum*. Therefore the mating system is a characteristic of a population and not of a species. In other taxa, different populations of the same species have sometimes different breeding systems (e. g. *Leavenworthia crassa*, Lloyd 1965; *Lycopersicon pimpinellifolium*, Rick *et al.* 1977; *Lycopersicon hirsutum*, Rick *et al.* 1979; *Gilia achilleifolia*, Schoen 1982; *Eichhornia paniculata*, Glover and Barret 1986; *Turnera ulmifolia* Barrett and Shore 1987; *Clarkia temblorіensis*, Holtsford and Ellstrand 1989; *Mimulus guttatus*, Ritland and Ritland 1989). Thus in studies to test the effect of the breeding system on levels of genetic diversity, the mating system must be determined.

None of the four morphometric variables examined here seems to be significantly correlated with selfing rates ($P > 0.05$). However two of the four variables analysed were found to be very plastic. With our sample sizes, this may preclude finding a significant association between these characteristics and the mating system, even if there is one.

Therefore none of the four morphometric characteristics here examined can be used as a indicator of selfing rates in natural populations and closely related species of the *Antirrhinum* and *Misopates* genera. Contrary to theoretical expectations (see Results), the between-family component of variance (for all morphometric characters) is greater than the within family component for both self-compatible and self-incompatible populations. If the assumption of random mating is violated in the self-incompatible populations this pattern could be obtained.

Chapter 4

Evolution of the cycloidea gene family in *Antirrhinum* and *Misopates*

4.1. Introduction

At least six genes are known to be involved in development of the zygomorphic flowers in *Antirrhinum* (Reeves and Olmstead 1998), yet little is known about how these genes evolve. While many development genes (like the MADS-box genes) may be pleiotropic (Kramer *et al.* 1998), genetic analyses of *Antirrhinum* mutant lines suggest that cycloidea (*cyc*) is involved only in the shaping of the flower (Luo *et al.* 1996). This gene has an uninterrupted open reading frame encoding a putative protein of 286 amino acids that is necessary to establish full dorsoventral asymmetry (Luo *et al.* 1996). Mutant lines of *Antirrhinum majus* in which all petals and stamens resemble their ventral counterparts in wild type (peloric phenotype) revealed that another gene, dichotoma (*dich*), is also necessary for full dorsoventral asymmetry (Luo *et al.* 1996; Almeida *et al.* 1997). This gene has recently been sequenced in *Antirrhinum majus*. It has an uninterrupted open reading frame encoding a putative protein of 314 amino acids that is 66% identical to *cyc* (Luo *et al.* 1999). Thus these two genes seem to be the result of an

ancient gene duplication event. Although they retain a common role in dorsal asymmetry and are partially redundant in the control of flower asymmetry, they have diverged with respect to particular functions; *cyc* affects the development of all petals and the *cyc* mutants often show an increase of organs per whorl, while *dich* is restricted to petals in dorsal positions and has little effect on flower organ number (Luo *et al.* 1999).

Both *cyc* and *dich* genes have a putative nuclear localisation signal, suggesting that they may play a role in transcription regulation (Luo *et al.* 1996; 1999). The *cyc* and *dich* (as revealed by database searches) proteins share one conserved region (by structural criteria this is predicted to be a basic-Helix-Loop-Helix, the bHLH domain) with other plant proteins, including the maize teosinte branched 1 (*TB*1), the rice *PCF*s genes, and several sequences of unknown function from *Arabidopsis thaliana* (Fig. 4.1; Doebley *et al.* 1997; Cubas *et al.* 1999a). Therefore these genes belong to a large family, named the "TCP domain" (based on the initial letters of the founding members *TB*1, *CYC*, *PCF*) that are probably involved in DNA- binding and protein interactions, according to the evidence that the bHLH domain of the *PCF* proteins is sufficient for DNA binding and necessary for dimerisation (Kosugi and Ohashi 1997; Cubas *et al.* 1999a). The presence of a second conserved region (rich in polar residues such as arginine, lysine and glutamic acid, named the R-domain) in *cyc*, *dich*, *TB1* and some *Arabidopsis* sequences, together with differences in the TCP domain, further define the *cyc*/*TB*1 subfamily (Fig. 4.1; Cubas *et al.* 1999a). The homologue of *Antirrhinum cyc* in *Linaria vulgaris* (L*cyc*) has been recently isolated and implicated in the control of floral symmetry. This gene is extensively methylated and transcriptionally silent in the peloric *Linaria* mutant (Cubas *et al.* 1999b).

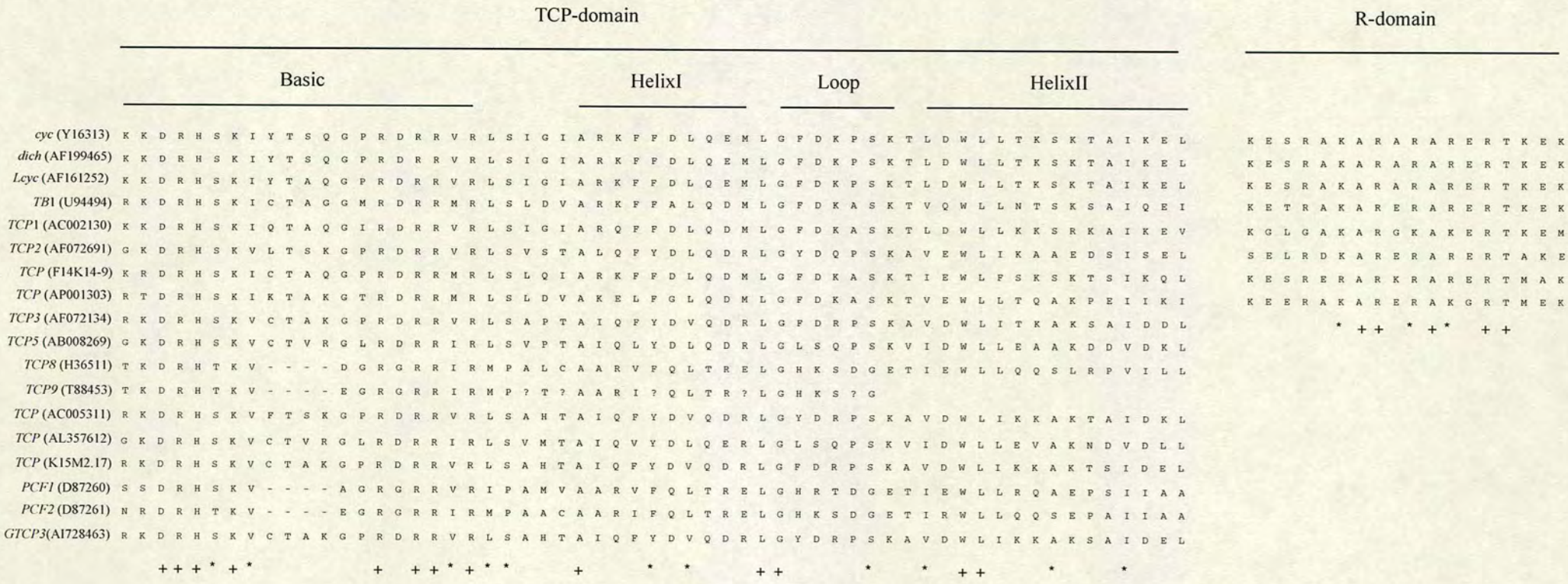


Fig. 4.1. Amino acid alignment of the TCP- and R -domain of TCP gene family members (modified from Cubas *et al.* 1999a). Sequences are from *Antirrhinum majus* (*cyc* and *dich*), *Linaria vulgaris* (*Lcyc*), maize *(TB*1), *Arabidopsis thaliana* (*TCP*s), and *Gossypium (GCP).*+ and * are conserved and conservative (defined as in Dayhoff *et al.* 1979) amino acid replacements, respectively; ? is an undetermined amino acid; dashes indicate deletions.

Members of the TCP family have also been described in the family Gesneriaceae (G*cyc*; Möller *et al.* 1999; Citerne *et al.* 2000), a family where most genera possess zygomorphic flowers. Although G*cyc* was believed to be a single copy gene in this family (Doyle and Doyle 1999, Möller *et al.* 1999), several paralogues have recently been described (Citerne *et al.* 2000). Phylogenetic analysis of G*cyc* genes in the family Gesneriaceae suggests that independent duplication and gene loss events have occurred during the evolution of this family after the split from Scrophulariaceae (Citerne *et al.* 2000).

In the course of a study attempting to estimate DNA sequence diversity for nuclear loci within species of *Antirrhinum* and *Misopates*, I found evidence for the existence of at least six functional genes (the *cyc* gene already known which I will call *cyc1A*, see below, plus *cyc1B*, *cyc2*, *cyc3, cyc4A*, and *cyc4B*), all of which are members of the TCP gene family. Although ancient duplications were already predicted since *cyc* and *dich* genes share sequence homology, this work revealed the presence of several recent duplications that made the study of diversity very difficult. *cyc1A*, *cyc1B*, *cyc2* and *cyc4A* genes are present in *Antirrhinum, Misopates* and *Digitalis*, suggesting that these gene duplications occurred before the split leading to *Digitalis* and *Antirrhinum / Misopates.*

4.2. Materials and Methods

Plant material

Leaves from all the species except *Antirrhinum valentinum*, *A. siculum*, and the Alg population of *A. majus* subsp. *cirrigherum* were collected in the field in summer

1997 and 1998 (Table 4.1). Leaves from *A. valentinum* and *A. siculum* were kindly provided by Isabel Mateu, while leaves from *A. majus* subsp. *cirrigherum* (population Alg) were kindly provided by Andrew Hudson.

DNA extraction and PCR amplification

Genomic DNA was prepared from leaves of individual plants using the method of Ingram *et al.* (1997). The *cyc* gene sequence deposited in GenBank (accession number Y16313) was used to design the primer pairs (all positions are relative to the *cyc* start codon): P1 5' TTGGGAAGAACACATACCTA 3' (position 5) and P2 5' AATTGATGAACTTGTGCTGAT 3' (position 859); Cu 5' CTTGAGTCCACCGCTTTGTT 3' (position 154) and Cl 5' CGTTGCCATAGTTTTGCTGA 3' (position 772). The primers C1 5' ACCACCACGGCCACCACCA 3' and C2 5' AAATCCAAACATTGAAGGG 3' were designed based on our determination of the sequence of *cyc4* (see below); they specifically amplifies *cyc3* and *cyc4* but not *cyc1A*, *cyc1B* or *cyc2* (see below). Based on the GenBank sequence, PCR reactions using primers P1 and P2 are expected to amplify an 855-bp fragment; Cu and Cl are expected to yield a 619-bp fragment (the latter set of primers was used to amplify the genes *cyc1A*, *cyc1B*, and *cyc2*; see below). Based on the new *cyc3* / *cyc4* genes (see below) C1 and C2 are expected to yield a 749-bp fragment. Standard amplification conditions were 30 cycles of denaturation at 94° C for 30 s, primer annealing at 53° C for 30 s, and primer extension at 72° C for 2 min.

Table 4.1 Taxa, code numbers, collection localities and years.

Species	Population code	Locality	Collection year
Misopates			
orontium	orontiumB	Portugal-Braganca	1997
	orontiumG	Portugal-Vila Nova de Gaia	1997
calycinum	calycinum	Portugal-Coimbra	1998
Antirrhinum			
molle lopesianum	molle	Portugal-Braganca	1998
braun-blanquetii	braun-blanquetii	Potugal-Braganca	1998
graniticum	graniticumB	Portugal-Braganca	1997
	graniticumS	Portugal-Chaves	1998
majus subsp. *linkianum*	linkianum	Portugal-Coimbra	1998
majus subsp. *cirrhigerum*	cirrhigerumAve	Portugal-Aveiro	1997
	cirrhigerumAlg	Portugal-Algarve	
valentinum	valentinum		
siculum	siculum2	Italy-Sicily	
	siculum6	Italy-Napoly	
	siculum7	Israel	
Linaria			
triornithophora	Linaria	Portugal-Braganca	1998
Cymbalaria			
muralis	Cymbalaria	U. K.-Edinburgh	1998
Digitalis			
purpurea	Digitalis	Portugal-Braganca	1998

Cloning and sequencing

The PCR products were cloned into the pCR 2.1 vector, using the TA cloning kit (Invitrogen). DNA sequencing was performed with an Applied Biosystems model 377 DNA sequencing system with the ABI PRISM Dye Termination cycle-sequencing Kit (Perkin Elmer), using the primers for the M13 forward and M13 reverse priming sites of the pCR 2.1 vector. Since a low rate of nucleotide mis-incorporation occurs in PCR reactions, it is known that this approach will lead to some sequencing errors. Such errors may be a problem for detailed diversity studies, but are usually negligible when one is characterising gene duplications or estimating divergence, since they represent a small proportion of the number of sequence differences. Some of the newly characterised cycloidea genes are very similar (see Results and Discussion), however, and therefore it is technically very difficult to detect possible sequencing errors.

Analyses of the sequences

DNA sequences were deposited in GenBank (accession numbers AF146833- 146848 for *cyc1*A; AF146849- 146862 for *cyc1B*; AF146863 - 146871 for *cyc2*; AF146872- 146873 for *cyc3*; and AF146874- 146880 for *cyc4*). The nucleotide sequences to be compared were aligned using ClustalX v. 1.64b (Thompson *et al.* 1997), and minor manual adjustments were performed using SeqPup v. 0.6f (Gilbert 1995). The numbers of synonymous and nonsynonynous differences between pairs of sequences were calculated using the DnaSP software (Rozas and Rozas 1997). Neighbor-joining trees were generated using the homologous regions sequenced in all the genes, with MEGA v. 1.01 (Kumar *et al.* 1994).

4.3. Results

4.3.1. Characterisation of Antirrhinum TCP loci

(i) Evidence for a new cycloidea locus, *cyc2*

Although the primers P1 and P2 (see Material and Methods) were designed based only on the *cyc* sequence (since the *dich* sequence was not available at the time this work was done) the P1 priming site is not present in the *dich* sequence (see below). When PCR was performed using these primers two bands with different sizes were observed. One is 855-bp (the expected fragment size; see Materials and Methods) and the other 950-bp. The 950 bp band is not expected to be the amplification product of *dich*.

The smaller band was cloned from two individuals of all taxa analysed (Table 4.1), and in 30 different plants the nucleotide sequences were determined. When these sequences were analysed, it was evident that, while some sequences from *Misopates orontium* were only 840-bp, all the others were 855-bp long. This 15-bp difference cannot be resolved in a 1.6 % agarose gel. Blast search reveals that both bands (840-bp and 855-bp) have high nucleotide similarity with *cyc* (> 92% acc. number Y16313). These two bands share amino acid similarity with L*cyc* (acc. number AF161252) and *dich* (acc. number AF199465), mainly due to the presence of the TCP and R -domains.

Among the 30 DNA sequences analysed there were 103 variable nucleotide sites and four variable indels. The sequences with the shorter and longer amplification products differ by 40 fixed differences and three indels. The large number of differences strongly suggests that these two groups of sequences represent two genes. I therefore tested whether both types of sequences are present in every individual analysed.

Between the two groups of sequences there is a fixed difference for a *Msp*I restriction site (present in the 840-bp group of sequences but not in the 855-bp sequences). We used a shorter amplification product (619-bp, using the Cu and Cl primers; see Materials and Methods) that still includes the region containing the *Msp*I site, to test for the presence of the sequence types. Figure 4.2 shows the restriction patterns of the products obtained from two randomly chosen individuals of each species (note that four individuals of each species were analysed, but only two of each species are shown in the figure because no differences in the banding patterns were seen). These results show that both types of sequences are present in every individual of every species analysed, i.e. I see the pattern expected if there are two different genes. The *cyc* gene deposited in GenBank does not have the *Msp*I restriction site, and I therefore named the gene with the *Msp*I restriction site *cyc2*. The fixed *Msp*I restriction site difference between the genes also allows us to clone and sequence the *cyc2* gene.

(ii) Evidence for a further cycloidea locus *cyc1B*

As shown above, the expected 855-bp band obtained with the P1 and P2 primers is not homogeneous, but is a mixture of two bands with different sizes (855-bp and 840-bp). When the sequence data from the 855-bp band were analysed, the complete set of sequences could again be classified into two groups, differing by three nucleotides (at positions 253, and 254, group 1 has G and G respectively, and group 2 has A and A, while at position 539, group 1 has C and group 2 has G). In addition, a further variant was found between the two groups at position 786 (A in most group 1 sequences, and G in group 2). This difference is not completely fixed, because the sequence from the Alg

population of *A. m.* subsp. *cirrhigerum* belongs to group 2 based on positions 253, 254 and 539, but has an A at position 786. These sequences could thus be alleles of the same gene, or could represent a recent duplication.

The two nucleotide differences at positions 253 and 254 create an additional *Aci*I restriction site in group 1 that is absent in group 2; this can be used to test whether these two groups of sequences are different alleles or different genes, as in the *cyc1* / *cyc2* analysis. For this test, once again, the shorter amplification product using the Cu and Cl primers was used. Unfortunately, the restriction pattern of the amplification products obtained from two randomly chosen individuals of each species was not conclusive. I could not show the presence of the *Aci*I site in some of our positive controls (genomic DNA from individuals that were known from our previous sequencing data to possess the *Aci*I restriction site); this is probably because of unequal amplification of the two types of sequences. However, I had previously cloned the 855-bp band (obtained with the P1 and P2 primers) from one individual of each species, and I therefore screened several clones to check if both types of sequences were present in every individual. The screening procedure involved amplifying 619-bp products from clones and digesting with *Aci*I. These tests clearly established that both types of sequences are amplified from every individual from every species analysed (Figure 4.3), the expected pattern if these are two different genes. I denote by *cyc1A* the gene that (like the *cyc* gene deposited in GenBank) does not have the *Aci*I restriction site, while the gene with the *Aci*I restriction site was named *cyc1B*.

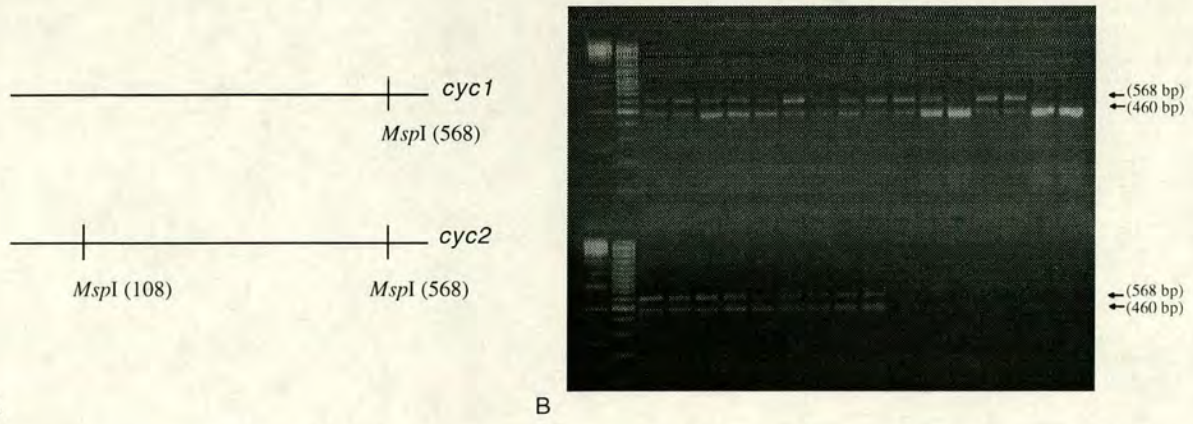


Fig.4.2. - Evidence that the *cyc1* and *cyc2* genes are present in every individual of each species analysed. (A) Schematic representation of the *Msp*I restriction sites present in the 619-bp amplification product (obtained with the Cu and Cl primers) of the *cyc1* and *cyc2* genes. The fixed difference in the *Msp*I restriction site between *cyc1* and *cyc2* can be used to distinguish the two genes. (B) *Msp*I restriction pattern of the the 619-bp amplification product of *cyc1* and *cyc2*. The arrows point to the expected 568-bp and 460-bp bands that are always observed; the 51-bp and 108-bp bands are not possible to resolve in a 1.6% agarose gel. Sequences names are according to the population code in Table 4.1. From left to right at the top: 1 Kb and 100 bp DNA ladder; orontiumB -1 and -2; orontiumG -1 and -2; calycinum -1 and -2; molle -1 and -2; braun-blanquetii -1 and -2; cirrhigerumAlg -1 and -2; cirrhigerumAve -2 and -3; linkianum -7 and -9; from left to right at the bottom: 1 Kb and 100 bp DNA ladder; graniticumB -1 and -2; graniticumS -1 and -2; valentinum -241; siculum6 -13; siculum7 -2; Digitalis; Cymbalaria; and Linaria.

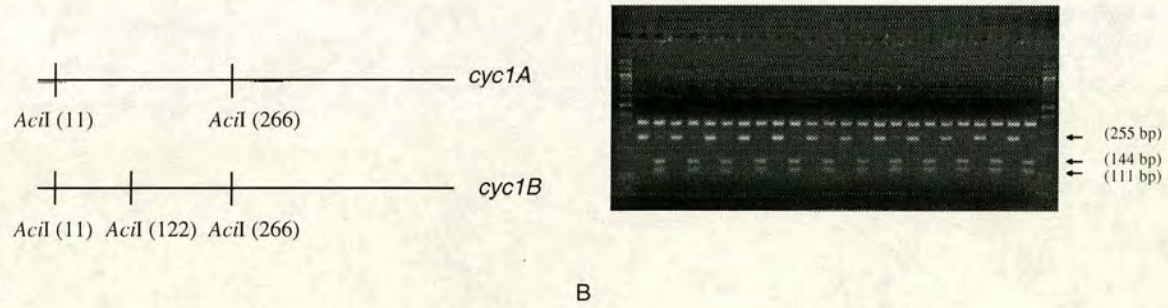


Fig. 4.3.- Evidence that the *cyc1A* and *cyc1B* genes are present in every species analysed. (A) schematic representation of the *Aci*I restriction sites present in the 619-bp amplification product (using the Cu and Cl primers) of the *cyc1A* and *cyc1B* genes. The fixed difference in the *Aci*I restriction site between *cyc1A* and *cyc1B* can be used to distinguish the two genes. (B) *Aci*I restriction pattern of the 619-bp amplification products. Each pair of lanes is from the same individual (different colonies) and show the presence of the two genes, *cyc1A* and *cyc1B*. These colonies were obtained after cloning an apparently homogeneous PCR product (855-bp band, obtained with the P1 and P2 primers) from one individual of each species. The arrows point to the expected 255-bp, 144-bp, and 111-bp bands that allow the two genes to be distinguished. Sequences names are according to the population code in Table 4.1. From left to right at the top: 100 bp DNA ladder; lanes 2 and 3 - orontiumB 5; lanes 4 and 5 - calycinum 2; lanes 6 and 7 - molle 1; lanes 8 and 9 - braun-blanquetii 1; lanes 10 and 11 - cirrhigerumAlg 1; lanes 12 and 13 - linkianum 7; lanes 14 and 15 graniticumB 1; lanes 16 and 17 - valentinum 241; lanes 18 and 19 - siculum7 2; lanes 20 and 21 - Digitalis; lanes 22 and 23 - Cymbalaria ; lanes 24 and 25 - Linaria; lane 26 - 100 bp DNA ladder.

(iii) Evidence for a further cycloidea locus *cyc4*

Two bands with different sizes are observed on 1.6% agarose gels, when PCR is performed using primers P1 and P2 (see above). Besides the 855-bp band (the expected fragment size; see Materials and Methods), an additional 950-bp amplification product is also observed. When the primers P1 and P2 were designed the *dich* sequence was not available, although recently it has been published. The P1 primer sequence is not conserved in this gene sequence, so this primer combination should not support the amplification of *dich.* To test whether the 950 bp band is a further TCP member, we therefore cloned and sequenced this amplification product from two individuals from both *A. m.* subsp. *cirrhigerum* and *A. graniticum*. Similarity with *cyc* is observed mainly due to the presence of the TCP and R -domains. Therefore it is likely that the 950 bp sequences represent another TCP-related gene, which I named *cyc4*.

When the *A. m.* subsp. *cirrhigerum* and *A. graniticum* sequences are compared, 18 nucleotide variable sites plus one indel (three bp long) are revealed. Nine of these variable sites are singletons and therefore may represent nucleotide mis-incorporation during amplification in the PCR. Four out of the nine variable sites that are not singletons are nonsynonymous. Although the sample size is small, these nine variable sites seem to be fixed differences between the *A. m.* subsp. *cirrhigerum* and *A. graniticum* sequences. If these *cyc4* sequences are orthologous, I expect divergence between *Digitalis purpurea*, *Misopates orontium* and either *A. m.* subsp. *cirrhigerum* or *A. graniticum* to be larger than between the two *Antirrhinum* species.

The 746 bp amplification product with the C1 and C2 primers (which do not amplify *cyc1A*, *cyc1B,* or *cyc2)* was cloned and sequenced in four individuals of *M. orontium* and one individual of *D. purpurea*. Two of the *M. orontium* sequences and the *D. purpurea* sequences were identical to that of *A. m.* subsp. *cirrhigerum cyc4*. This strongly suggests that *A. m.* subsp. *cirrhigerum* and *A. graniticum cyc4* sequences are not orthologous. I tentatively use the name *cyc4A* for the gene obtained from *A. m.* subsp. *cirrhigerum, M. orontium and D. purpurea,* and *cyc4B* for the gene obtained from *A. graniticum*.

Both *cyc4* genes share high nucleotide similarity with another TCP gene from *Antirrhinum majus*, *dich* (>97% nucleotide identity), that has recently become available in GenBank. However between *cyc4A* and *dich* there are eight synonymous, six non-synonymous differences, and two indels (12 bp and three bp long), and between *cyc4B* and *dich* there are four synonymous, eight non-synonymous differences, and one indel (12 bp long). Therefore both *cyc4A* and *cyc4B* are *dich*-related genes, but are unlikely to be orthologues of *dich*.

(iv) Evidence for *cyc3*

The amplification product using primers C1 and C2 was cloned and sequenced in four individuals of *M. orontium*. Two of these sequences were identical to *A. m.* subsp. *cirrhigerum cyc4* (see above). When the other two sequences were compared with the *cyc4* genes, the comparison revealed 52 fixed nucleotide differences (36 synonymous and

16 non-synonymous) plus two indels (three bp long). The large number of differences strongly suggests the presence of another *dich*-related gene. I named this gene *cyc3*. Blast search of the *cyc3* sequences revealed high nucleotide sequence similarity (90% nucleotide identity) with *dich*. Therefore *cyc3* is a *dich*-related gene, but is unlikely to be an orthologue of *dich*.

4.3.2. DNA variability within and between species

Sequence variation was estimated for four of the genes described above, both within and between species (see Tables 4.2, 4.3, 4.4, and 4.5), and in terms of the average numbers of synonymous differences per synonymous site (K_s) and of non-synonymous differences per non-synonymous site (K_a).

Table 4.2. *cycIA* sequence variation within and between species. Mean pairwise numbers of non-synonymous differences per non-synonymous site, K_a (above the diagonal) and synonymous differences per synonymous site, K_s (below the diagonal). All values are multiplied by 10^3. The sequences compared are 814-bp long. When more than one sequence from a species is included, the number of sequences used is shown (N).

cycIA	*A. m.*subsp. *majus*	*A. m.* subsp. *cirrhigerum* N=2	*A. m.* subsp. *linkianum* N=3	*A. graniticum* N=4	*A. braun-blanquetii*	*A. molle*	*A. valentinum*	*A. siculum* N=2	*M. orontium*	*Digitalis*
A. m. subsp. *majus*	----------	4.4	2.3	3.9	3.5	1.7	7.0	3.5	3.5	5.3
A. m. .subsp. *cirrhigerum*	34.0	5.3 43.4	6.8	6.6	4.4	2.6	7.9	7.9	4.4	6.2
A. m. subsps. *linkianum*	39.9	29.7	4.7 8.1	6.3	5.8	4.1	9.4	5.8	5.8	7.6
A. graniticum	23	29.2	17.7	7.4 19.4	5.7	4.0	9.3	7.5	5.6	7.5
A. braun-blanquetii	30.8	34.0	24.5	23.0	----------	1.7	7.0	5.3	0.0	5.3
A. molle	24.5	27.7	18.2	16.8	6.1	----------	5.3	5.3	1.7	3.5
A. valentinum	30.7	40.2	24.5	19.9	24.5	18.3	----------	10.6	7.0	8.8
A. siculum	34.0	37.1	15.2	23.0	27.7	21.4	27.6	0.0 6.1	5.3	8.8
M. orontium	30.8	34.0	24.5	23.0	0.0	6.1	24.5	27.7	----------	5.3
Digitalis	30.8	40.3	24.5	23.0	24.5	18.3	24.5	27.7	24.5	----------

Table 4.3. *cyc1B* sequence variation within and between species. See Table 4.2 for details. The sequences being compared are 814-bp long.

cyc1B	*A. m.* subsp. *cirrhigerum* N=2	*A. braun-blanquetii*	*A. molle*	*A. valentinum* N=2	*A. siculum* N=2	*M. orontium*	*M. calycinum* N=2	*Digitalis*	*L. triornithophora* or *C. muralis*
A. m. subsp. *cirrhigerum*	0.0 0.0	0.0	1.7	3.4	0.9	1.7	3.4	5.1	0.0
A. braun-blanquetii	5.8	_____	1.7	3.4	0.9	1.7	3.4	5.1	0.0
A. molle	0.0	5.8	_____	5.1	2.6	3.4	5.1	6.8	1.7
A. valentinum	14.7	20.7	14.7	3.4 29.7	4.3	5.1	5.1	8.5	3.4
A. siculum	0.0	5.8	0.0	14.7	1.7 0.0	2.6	4.3	6.0	0.9
M. orontium	0.0	5.8	0.0	14.7	0.0	_____	5.1	6.8	1.7
M. calycinum	2.9	8.8	2.9	17.7	2.9	2.9	6.8 5.8	8.5	3.4
Digitalis	5.8	11.7	5.8	20.8	5.8	5.8	8.8	_____	5.1
L. triornithophora or *C. muralis*	0.0	5.8	0.0	14.7	0.0	0.0	2.9	5.8	_____

Table 4.4. *cyc2* sequence variation within and between species. See Table 4.2 for details. The sequences being compared are 462-bp long.

cyc2	*A. m.* subsp. *cirrhigerum* N=2	*A. graniticum* N=2	*M. orontium* N=4	*Digitalis*
A. m. subsp. *cirrhigerum*	0.0 20.1	2.9	2.1	5.6
A. graniticum	15.1	5.7 10.0	5.0	8.5
M. orontium	12.6	10.0	4.3 10.0	7.9
Digitalis	10.0	15.1	10.1	_____

Table 4.5. *cyc4A* sequence variation within and between species. See Table 4.2 for details. The sequences being compared are 711-bp long.

cyc4A	*A. m.* subsp. *cirrhigerum* N=2	*M. orontium* N=2	*Digitalis*
A. m. subsp. *cirrhigerum*	7.5 19.9	3.7	3.8
M. orontium	10.0	0.0 0.0	0.0
Digitalis	10.0	0.0	_____

For the *cyc1A*, *cyc2,* and *cyc4* genes, all K_a / K_s ratios, both within and between species, are less than one (Table 4.6). These ratios suggest that purifying selection is acting on these genes. For *cyc1B*, however, both within or between populations, K_a / K_s

ratios higher than one are obtained for approximately 50% of the comparisons. Although 22 nucleotide positions differ among the 14 *cyc1B* sequences, 21 of these variable sites are singletons. The unique non-singleton nucleotide difference among the 14 *cyc1B* sequences is a non-synonymous replacement site. Therefore, there is no evidence to support the view that *cyc1B* sequences are different genes.

Table 4.6. Proportions of K_a / K_s ratios higher than 1, within and between species, for the *cyc1A*, *cyc1B*, *cyc2* and *cyc4* genes.

	Within species			Between species		
	Number of comparisons	K_a / K_s >1 *	Mean K_a / K_s ratio §	Number of comparisons	K_a / K_s >1 *	Mean K_a / K_s ratio §
cyc1A	4	0	0.27	45	0	0.23
cyc1B	4	2	-	36	17	-
cyc2	3	0	0.33	6	0	0.46
cyc4A	2	0	0.38	3	0	0.32

* when more than one sequence was available for the same species, the average of the K_a / K_s ratio was used; § the mean K_a / K_s ratios are the averages of the ratios that could be calculated, i.e. only the cases where K_s is non-zero. Altogether, for *cyc1A*, *cyc2* and *cyc4*, there were only two such cases between species, and two within species. For *cyc1B*, K_a / K_s ratios were not calculated, because of heterogeneity of the values.

4.3.3. Genealogical relationships among the Scrophulariaceae TCP genes

The relationships among the different Scrophulariaceae TCP- genes are shown in the gene tree in Figure 4.4. The sequences for each putative gene group together, and bootstrap values are high, supporting our interpretation that the different groups of sequences correspond to different genes. Because four of the five Scrophulariaceae genes analysed (*cyc1A*, *cyc1B*, *cyc2,* and *cyc4A*), are also found in *Digitalis,* these gene duplications must have occurred before the split between the lineages leading to *Digitalis* and *Antirrhinum*. No phylogenetic information based on molecular markers is available for *Cymbalaria* and *Linaria*. However, based on morphological data, *Cymbalaria* and *Linaria* species seem to be more closely related to *Antirrhinum* than to *Digitalis* (the first description of the genus *Antirrhinum* by Linnaeus in 1753 (cited in Sutton 1988), included species of *Antirrhinum*, *Linaria* and *Cymbalaria*). Here I show that *cyc1A*, *cyc1B*, *cyc2,* and *cyc4* are present in *Cymbalaria muralis* and *Linaria triornithophora* (Fig. 4.4). The *Linaria vulgaris Lcyc* gene has been reported to be the homologue of *Antirrhinum cyc*, but based on the K_s values between this gene and the *Antirrhinum/ Misopates* TCP genes, seems not to be the orthologue of any of the TCP genes described here (Fig. 4.4; the K_s values, after Jukes-Cantor correction, between L*cyc* and *cyc1A*, *cyc1B*, *cyc2*, *cyc3*, and *cyc4* are, respectively, 0.77, 0.75, 0.67, 1.17 and 1.12).

The time of the divergence of *Digitalis* and *Antirrhinum* / *Misopates* can be roughly estimated to be 5 Myr, if I assume that Scrophulariaceae and Solanaceae have been separated for 40 Myr (Xue *et al.* 1996; see Discussion in Chapter 6) and that branch lengths in Figure 15B in Chase *et al.* (1993; see Appendix) are linearly related to time, which is, of course, dubious. With this dating, the synonymous mutation rates for

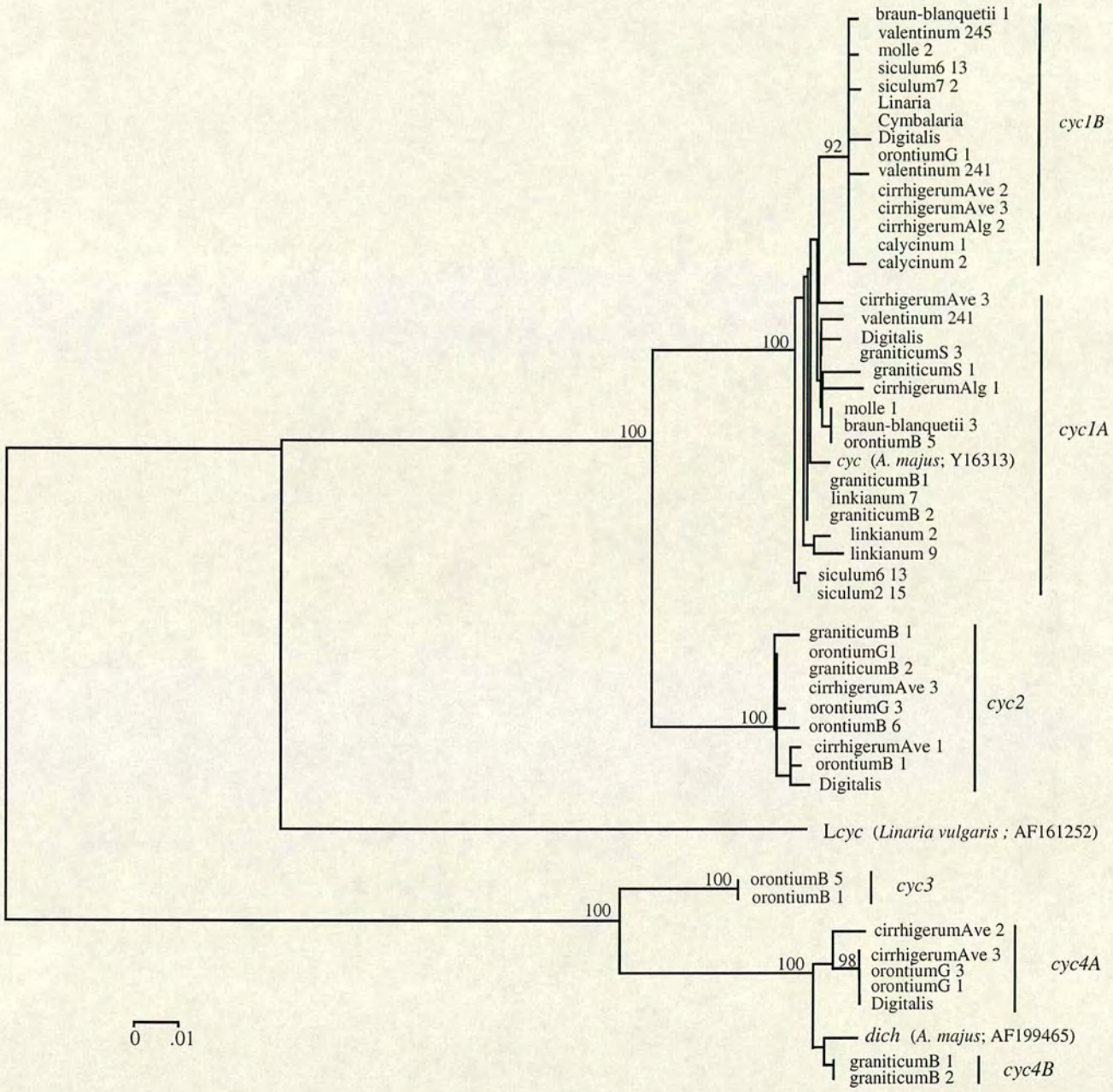


Fig. 4.4. Consensus neighbor-joining tree using Jukes-Cantor correction, showing the relationship among the different *cyc* related DNA sequences that have been studied (only the 468 bp sequence that is common to all the DNA sequences was used). Sequence names are according to Table 4.1. The numbers after the sequences names are sample codes. The length of the terminal branches are not proportional to the distance values, because branch lengths in the tree are proportional to the numbers of inferred nucleotide site changes according to the reconstruction rather than to the distance values. Percentages of bootstrap replicates supporting the branches are shown where the values exceed 75%.

these loci can be estimated, by dividing the K_s values between all *Antirrhinum* / *Misopates* sequences and the putative *Digitalis purpurea* sequences of the same gene by 10 Myr (twice the estimated divergence time). The mutation rates per year were estimated as the average of these values, and are as follows: *cyc1A* 2.7 x 10^{-9}; *cyc1B* 1.0 x 10^{-9}; *cyc2* 1.1 x 10^{-9}. However these mutation rates may be overestimates, since other phylogenies (Reeves and Olmstead 1998; Wolfe and dePamphilis 1998) have suggested an older date for the split between *Digitalis* and *Antirrhinum* / *Misopates*.

4.4. Discussion

The data presented here are compatible with the presence of at least six TCP genes in *Antirrhinum*. There is no reason to believe that any of the TCP genes described in this work is a pseudogene. I have no direct evidence for expression, but none of the genes has any interruption of the open reading frame, and both within and between species the number of synonymous substitutions per synonymous site is usually much larger than the number of non-synonymous substitutions per non-synonymous site (Tables 4.2 - 4.6), as expected for genes under purifying selection.

Five duplication events, with different ages, must have occurred in the *Antirrhinum* lineage. Assuming a molecular clock (Zuckerkandl and Pauling 1965) estimates for the age of the events can be obtained by dividing the average *Ks* values between the different genes by the average mutation rate for all the *cyc* genes (1.6 x 10^{-9}). These estimates represent the minimum age of the duplications, since they assume 5 Myr for the split of *Digitalis* and *Antirrhinum* / *Misopates*. Using other estimates, the

duplication events would be much older. The oldest duplication event occurred between *cyc3 / cyc4A / cyc4B / dich* and *cyc2 / cyc1A / cyc1B*. I do not attempt to estimate the age of this event since the K_s values are at saturation. The duplication event between *cyc3 / cyc4* and also between *cyc2 / (cyc1A / cyc1B)* is then estimated to have occurred about 75 MYA. This date can be compared with the time of the origin of the angiosperms, estimated to have occurred 160 - 348 MYA (with considerable uncertainty; see Goremykin *et al.* 1997). However, no reliable angiosperm fossils older than 120 Myr have been found (Hickey and Doyle 1977; Doyle 1978). Therefore it is often assumed that the angiosperms began to radiate rapidly about 115 MYA (for further discussion see Li 1997).

These old duplications predate the split of several plant families. A history of gene duplications, concerted evolution among the duplicate genes and loss events would, however, make it very difficult to infer orthology among the TCP genes of different plant families. The relationships of the Scrophulariaceae, Gesneriaceae, maize and *Arabidopsis* TCP- genes, based on the TCP and R- domains are shown in Figure 4.5. The two domains comprise a small region of 129 bp and therefore it is unlikely that enough information exists to support the youngest gene duplications, namely *cyc1A* and *cyc1B* , *cyc4B* and *dich* in Scrophulariaceae and G*cyc1A* and G*cyc1B* in Gesneriaceae (Fig. 4.5). Multiple duplication events within each family group seem to have occurred. Although, the age of the split between Scrophulariaceae and Gesneriaceae is estimated to be only 20 Myr (based on palaeontological evidence; cited in Citerne 1999; Muller 1981) the TCP genes of these two families seem to have independent histories. However this may be the result of the incomplete characterisation of this gene family in both plant groups.

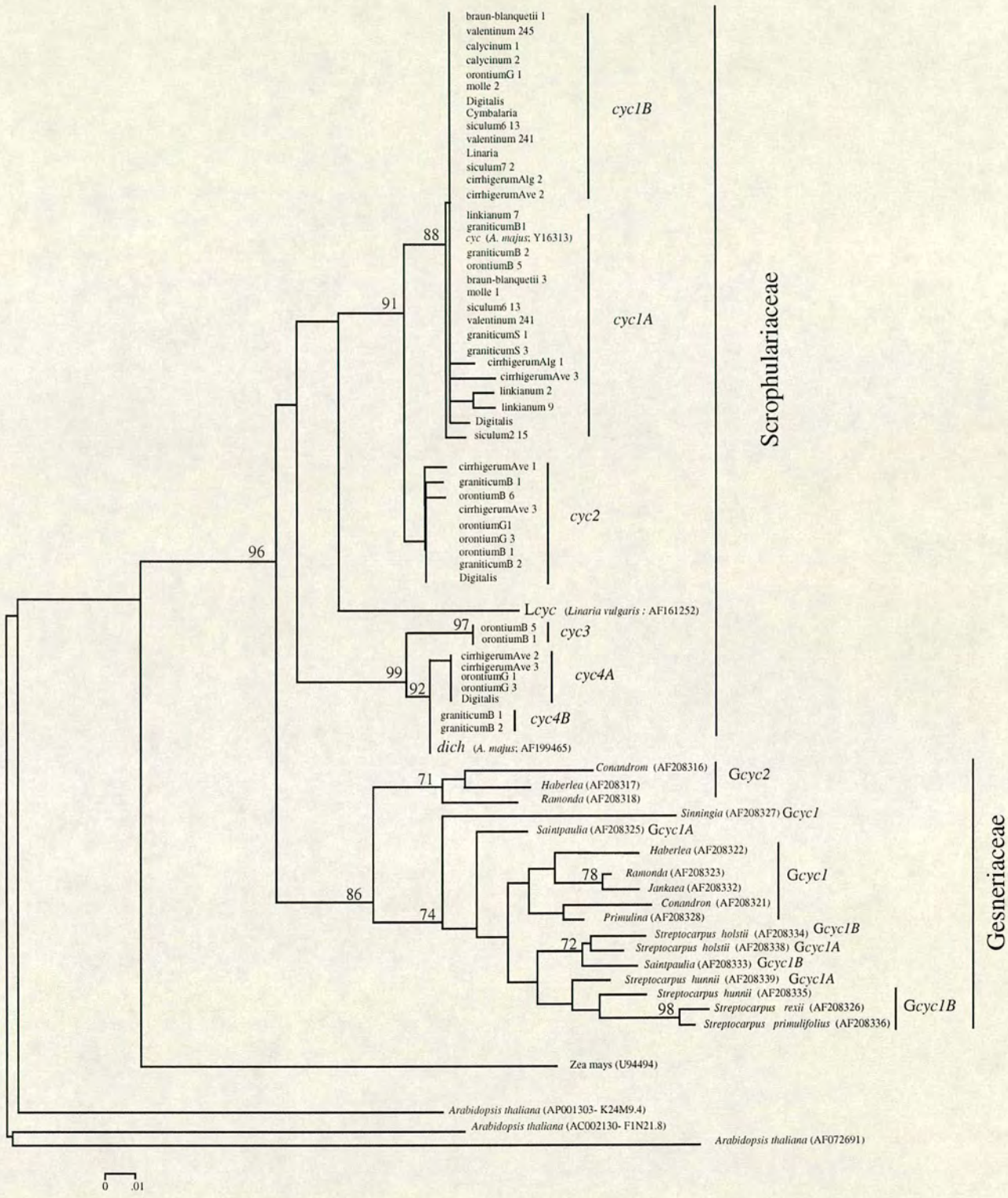


Fig. 4.5. Neighbor-joining tree using Kimura two-parameter distance, showing the relationship among the different TCP-related DNA sequences of Scrophulariaceae, Gesneriaceae, *Zea mays* and *Arabidopsis* based on the TCP and R- domains. The two domains comprise a small region of 129 bp. Percentages of bootstrap replicates supporting the branches are shown where the values exceeded 70%.

The estimated age of the *cyc1A* / *cyc1B* duplication event is 7.5 MYA, suggesting that this duplication occurred after the split between the Scrophulariaceae and Callitrichaceae on one hand, and the Acanthaceae on the other hand (based on information in Chase *et al.* 1993; see Appendix). No TCP genes have yet been characterised in these plant families.

Estimates of the age of the duplication events depend heavily on the mutation rate used. If the mutation rates are underestimates, the ages of the duplication events are overestimated. Our mutation rate estimate (1.6×10^{-9} ; the average of the rates for the different *cyc* genes) is lower than those obtained for nuclear genes in short-lived monocotyledonous plants ($5.1 - 7.1 \times 10^{-9}$; Wolfe *et al.* 1989; Gaut *et al.* 1996), and in palms (2.61×10^{-9}; Gaut *et al.* 1996), where it has been suggested that a generation-time effect may cause a reduced mutation rate. However, our estimate is similar to that obtained for the first published estimate for a dycotyledonous plant (*Gossypium sp.*; $1.47 - 2.05 \times 10^{-9}$; Small *et al.* 1999).

Having irregular flowers is a derived characteristic that seems to have appeared much longer ago than the *cyc1A* / *cyc1B* gene duplication, just before the split between most Scrophulariales and the Lamiales (Chase *et al.* 1993; see Appendix). The estimated 75 Myr of divergence that separates the *cyc4* / *cyc3* and the *cyc2* / (*cyc1A* and *cyc1B*) genes suggests that these duplications occurred before the diversification of the clade including Solanaceae, Cornaceae, Garriaceae, Ericaceae, and Hydrangeaceae (Chase *et al.* 1993). All these families have regular flowers (Heywood 1985). Therefore there seems to be no correlation between flower shape and these gene duplications.

The existence of this moderately sized gene family could very easily have escaped our notice, had I not examined diversity within the study species. Paralogous loci with very different sequences (such as *cyc1A*, *cyc1B*, and *cyc2* versus *cyc3* and *cyc4*) will become evident as soon as sequence data are collected, but sequences that are similar (such as *cyc1A* and *cyc1B,* and *cyc4A* and *cyc4B*) will often be assumed to be allelic variants of a single locus. It should be noted that within each gene sequence group several differences were observed, the majority of them being singletons. It is not known to what extent they represent nucleotide mis-incorporations during PCR amplification, since only one clone for each sequence was obtained. Different genes have very similar sequences, and therefore gene specific primers are difficult to design. Therefore direct sequencing is not possible. Many clones must be sequenced for each individual in order to eliminate PCR mis-incorporation, otherwise consensus sequences may be chimeras between very similar genes. TCP genes are thus unsuitable for detailed studies of sequence diversity, divergence, and phylogenetic analyses.

It is widely recognised that orthologous genes must be compared for phylogenetic inferences (e.g., Li 1997), as well as for the study of within-species diversity. Southern blotting studies are often used to provide evidence for orthology but definitive evidence is not always generated, because this approach relies on the assumption that sequences flanking the genes will differ. This may not always be the case, especially for very recent duplications as found in this study. At present, data on rates of accumulation of diversity in sequences flanking plant genes are few, apart from evidence for high diversity at maize and teosinte loci (Sanmiguel *et al.* 1997; Wang *et al.* 1999) and in the flanking regions of some self-incompatibility loci (Coleman and Kao 1992; Boyes *et al.*

1997). It is not widely realised how much work may be needed to obtain good evidence for orthology, nor is there widespread awareness that, perhaps unlike the situation in *Drosophila*, small gene families are of common occurrence among plant nuclear loci (e.g., Galloway *et al.* 1998).

Chapter 5

Low diversity and divergence in the *fil1* gene family of *Antirrhinum.*

5.1. Introduction

In the first attempt to estimate the level of DNA diversity in the genus *Antirrhinum*, and test for an effect of the breeding system on diversity, I have used the cycloidea (*cyc*) genes (Chapter 4). On average, the nucleotide site diversity (π, Nei 1987) was found to be 1.23%, with values ranging from 0 to 4.34% for different species. These diversity values were obtained using samples taken from several different populations of a species rather than from within single populations. Due to difficulties of defining orthologous genes in the TCP gene family, detailed sequence diversity studies could not be performed, and some of the highest π values were suspected to be caused by the presence of further unidentified TCP genes. The sequences obtained for several *cyc* loci were very little diverged between several *Antirrhinum* species and *Digitalis* (Chapter 4). The estimated nucleotide substitution rates for the *cyc* genes are lower than most other estimates for plant nuclear genes (based on several genes in monocotyledonous plants, see Gaut *et al.* 1996), but similar to an estimate from another dicotyledonous plant

(*Gossypium sp.*; Small *et al.* 1999). Since different *cyc* genes have probably been pooled, even these low nucleotide substitution rates are overestimates.

Several factors, including hybridisation among *Antirrhinum* species (Mather 1947; Harrison and Darby 1955; Herrmann 1973), as well as recent evolution of the species in this genus, could account for the low divergence. An unusually high level of constraint could also explain both the low diversity and divergence observed. In order to examine these issues more fully it is therefore desirable to obtain data on a single-copy gene, and to include regions expected to differ in constraint, such as introns and flanking regions. In plants, data on flanking sequences are currently very scanty and, in nuclear genes, variation in introns is often lower than silent variation in coding regions (reviewed by Charlesworth and Charlesworth 1998). This pattern is also observed in *Drosophila* where intron regions are, on average, less variable and diverge less than synonymous sites (Moriyama and Powell 1996; Bauer and Aquadro 1997).

The *fil1* gene has been cloned and sequenced in *Antirrhinum majus* (Nacken *et al.* 1991). It was described as a single-copy gene encoding a putative protein of 93 amino acids (Nacken *et al.* 1991) with one intron of 266 bp. This putative protein may be involved in cell wall formation of the cells of the filaments at the bases of the petals. Sequence data are available for 1.8 kb upstream of the transcription start site and 0.4 kb downstream of the polyadenylation site, so this gene seemed a good choice for DNA variability studies. I have therefore analysed the level of nucleotide variation in a 652 bp region that includes 38% of the coding region, the intron, and a portion of the 3' non-

coding region. Three geographically distinct populations of *Antirrhinum majus* subsp *cirrhigerum*, one population of *A. majus* subsp *linkianum*, and one population *of A. graniticum*, that have different breeding systems were studied.

We show here that *fil1* belongs to a gene family of at least five genes, which I named *fil1A1, fil1A2, fil1A3, fil1B,* and *fil1C*. Some members of this gene family are very similar and can differ in just one indel (in the intron) or a single nucleotide. I do not know whether these genes are arranged in tandem or are dispersed in the genome. Very low nucleotide diversity is observed for all the five genes, even in silent and intron sites. Low divergence is also observed for all *fil1* genes between *Digitalis* and *Antirrhinum*. For three of these genes I also obtained a sequence from *Verbascum nigrum*. Compared with *Antirrhinum*, there is very little divergence, the sequences being identical in two of the cases.

5.2. Material and Methods

Plant Material

Table 5.1 shows the sub-set of *Antirrhinum* populations and species I use in this Chapter. One individual of *Digitalis purpurea* was collected in the field in Portugal (Bragança; code digitalisB), and the other was a wild individual collected in the UK (Edinburgh; code digitalisE). *Verbascum nigrum* leaves were collected from a wild individual growing on the Edinburgh University campus.

Table 5.1. *Antirrhinum* species, code names and mating system.

Antirrhinum Species	Population code	Mating system
majus subsp. *cirrhigerum*	cirrhigerumAve	Self-compatible
	cirrhigerumGala	Largely self-compatible
	cirrhigerumMuel	Largely self- incompatible
majus subsp. *linkianum*	linkianum	Partially self- incompatible
graniticum	graniticumB	Self-incompatible

DNA Extraction and PCR Amplification

Genomic DNA was prepared from leaves of individual plants using the method of Ingram *et al.* (1997). The *fil1* gene sequence in GenBank (accession number X57296) was used to construct the following primers (all positions are relative to the *fil1* start codon): F1 (position 176) and F2 (position 809) (Table 5.2). Additional primers based on our new sequences for each of the *fil1* genes (see Results) were also used (Table 5.2). The primer combination AGAG (designed to amplify sequences with the insertion at position 341; Fig. 5.1) combined with F2 specifically amplifies *fil1A1*, AGAA (designed to amplify sequences without the insertion at position 341; Fig. 5.1) combined with F2 amplifies *fil1A2* and *fil1A3,* F1 plus BR (designed to amplify sequences with a G at position 630; Fig. 5.1) specifically amplifies *fil1B*, and CF (designed to amplify sequences with the insertion at position 283; Fig. 5.1) plus F2 specifically amplifies *fil1C*. Standard amplification conditions were 30 cycles of denaturation at 94° C for 30s, primer annealing at 51° C for 30 s, and primer extension at 72° C for 2 min. Negative

controls (PCR cocktail with no genomic DNA added) were included in all PCR amplifications and never yielded PCR products. Since only one *Verbascum nigrum* individual was used, DNA was extracted independently from several leaves and treated as different samples. The PCR of these samples always gave the same amplification products.

Table 5.2. Primers used in this study

Forward primers	
F1	5' CCGCCCTTCAGTCCGTAGA 3'
AGAG	5' CCTTACAAAAAACTTAGAG 3'
AGAA	5' CCTTACAAAAAACTTAGAA 3'
CF	5'GTGGTGGTGAGCCTAAACT 3'
Reverse primers	
F2	5' ATGTCTATTCTTTCTTGGA 3'
BR	5' AGTACATGATTCAGCACAC 3'

Cloning and Sequencing

The PCR products amplified using F1 and F2 were cloned into the pCR 2.1 vector, using the TA cloning kit (Invitrogen). DNA sequencing was performed with an Applied Biosystems model 377 DNA sequencing system with the ABI PRISM BigDye cycle-sequencing Kit (Perkin Elmer), using the primers for the M13 forward and M13 reverse priming sites of the pCR 2.1 vector. When specific primers were used, there is very low DNA variability in the region of the genes analysed (see Results), so DNA sequencing was performed directly from the PCR product. Very rarely, superimposed

ladders were observed, and in these cases the PCR products were cloned and sequenced as described above.

Analyses of the Sequences

The DNA sequences were deposited in GenBank (accession numbers AY007994 - 008028 for *fil1A1*, AY008029-008060 for *fil1A2,* AY008061-008064 for *fil1A3,* AY008065-008080 for *fil1B*, and AY008081-008086 for *fil1C*). The nucleotide sequences to be compared were aligned using ClustalX v. 1.64b (Thompson *et al.* 1997), and minor manual adjustments were performed using SeqPup v. 0.6f. The average number of synonymous and nonsynonynous differences between pairs of sequences (π; Nei 1987) were calculated using the DnaSP software (Rozas and Rozas 1997). For the analyses of codon usage bias I collected 52 sequences of *Antirrhinum* nuclear genes from GenBank (see Table 5.3 for accession numbers). I searched for the presence of potential conserved RNA helices conserved in the *fil1* genes using the program PIRANAH1.1 (Parsch *et al.* 2000). The program's default values were used. As will be seen, the *fil1* genes have similar sequences in the different species studied. I were therefore unable to use covariation between pairs of sites to detect regions of sequence likely to form helical structures. Instead, I used the alignment of the *fil1*A, *fil1B*, and *fil1C* genes (see below).

Table 5.3. ENC for 52 *Antirrhinum majus* genes.

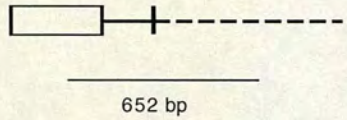
Gene	Acc. number	ENC	Gene	Acc. number	ENC
bZIP1	Y13675	32.62	*cyclin1*	X76122	54.60
ptl2	X71782	36.18	*DAG*	X95753	54.65
ubiquitin	X67957	38.69	*S2-Rnase*	X96465	55.06
DEFH49	X95467	46.01	*cyc4*	AF146875	55.25
GAPDH	X59517	47.50	*chs*	X03710	55.31
cdc2c	X97639	47.94	*fap3*	Y14858	55.60
globosa	X68831	48.46	*S4-Rnase*	X96466	55.63
squamosa	X63701	49.95	*cyc1B*	AF146849	55.90
bZIP2	Y13676	50.30	*S5-Rnase*	X96464	56.39
squamosa-promoter-binding-protein 2	X92079	50.49	*fim*	S71192	56.62
fap2	Y14857	50.53	*cdc2d*	X97640	56.77
DEFH125	Y10750	50.59	*ACS1*	AF083814	56.77
cdc2a	X97637	52.28	*mixta*	X79108	56.91
waxy	AJ006293	52.33	*hmr1*	AJ236702	57.21
DEFH72	X95468	52.44	*TAP1*	X57295	57.46
cdc2b	X97638	53.25	*cyc*	Y16313	57.53
DEFH200	X95469	53.35	*TAP2*	X55434	57.68
PHYB/D	U08143	53.43	*cyclin2*	X76123	57.86
magnesium chelatase	X73144	53.56	*plena*	S53900	58.58
ACS2	AF083815	53.61	*cyc2*	AF146868	58.97
Del	M84913	53.75	*cyc3*	AF146873	59.50
pallida	X15536	53.91	*fil2*	X76995	60.20
GTPase	Y17899	54.06	*PHYE*	U08144	60.31
flo	M55525	54.40	*SAP1*	AJ132349	60.48
deficiens	X52023	54.53	*phan*	AJ005586	60.63
DEFA	X62810	54.54	*ACS3*	AF083816	60.77

5.3. Results

5.3.1 fil1 is a Gene Family

Using primers F1 and F2, a PCR amplification product with the expected size (652 bp) was obtained. This was cloned and sequenced in several individuals of the three *A. majus* subsp *cirrigherum* populations (11 cirrhigerumAve, 11 cirrhigerumGala, 12 cirrhigerumMuel individuals) and three individuals of *A. majus* subsp *linkianum* (Table 5.4). The sequences cluster into three groups (A, B, and C) that differ by several nucleotide differences (Fig. 5.1; type A and B are similar in size and type C is longer). For these sequences D' (the scaled linkage disequilibrium between pairs of informative sites; Lewontin 1964) is always 1, which is unlikely for allelic variants (Miyashita and Langley 1988). Furthermore, it was possible to amplify all three types of sequences (A, B, and C) from every individual analysed (a minimum of five individuals of each *Antirrhinum* species; Table 5.4) using specific primers for each sequence type (see Material and Methods). This is incompatible with the *fil1* gene being single-copy. I will refer to these genes as *fil1A*, *fil1B* and *fil1C*.

These three genes were all found to be present in both *Digitalis* plants as well as *Antirrhinum*. The gene duplications that gave rise to this gene family must therefore pre-date the split between these genera. The *fil1* sequence in GenBank is most similar to *fil1A,* differing in only four nucleotides in the intron region that are fixed among the *fil1A* sequences (Fig. 5.1), but they are not necessarily the same locus. Note that the *fil1* GenBank sequence is a composite of the sequence of a cDNA clone (coding region) and a genomic clone (intron plus flanking regions; see Nacken *et al.* 1991), and therefore may



A

	Exon1		Intron																														3' flanking non-coding region											
	201	237	283	291	292	297	327	330	337	340	341	349	359	368	390	429	439	445	468	479	495	498	499	505	511	513	520	523	534	558	578	630	693	748	754	761	763	792	794	800				
GenBank sequence *A.majus*	G	G	-	A	T	T	T	C	-	C	-	G	C	G	C	C	A	A	T	T	T	T	A	-	T	T	G	T	-	A	A	C	T	A	G	T	T	T	G	-				
fil1A1	.	.	.	.	.	A	.	.	C	G	*i*(2)	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.				
fil1A2	.	.	.	.	.	A	.	.	C	G	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	T/C	.	.	.				
fil1A3	.	.	.	.	.	A	.	.	C	G	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.				
fil1B	.	.	.	.	.	.	.	.	C	.	.	C	.	.	T	.	G	.	.	.	.	.	.	.	.	.	T	.	.	.	.	G	.	C	.	A	.	A	T	*i*(3)				
	.	.	.	.	.	.	.	.	C	.	.	C	.	.	T	.	G	.	.	.	.	G	.	.	.	.	T	.	.	.	.	G	.	C	.	A	.	A	T	*i*(3)				
fil1C	A	A	*i*(17)	T	G	A	A	T	T	.	.	C	T	A	.	A	.	T	C	A	A	.	T	*i*(2)	A	A	.	C	*i*(2)	T	G	.	.	.	C	A	.	.	.	.				

B

Fig. 5.1. (A) Schematic diagram of *fil1* GenBank sequence. Boxes represent exons, the solid line the intron, and the dotted line the 3' non- flanking sequence (modified from Nacken *et al.* 1991). The 652 bp region analysed is indicated by a double head arrow. (B)Variable sites of *fil1* - like genes. Dots represent the same nucleotide as in the first sequence, dashes indicate deletions, *i* indicate insertions (the numbers in brackets represent the size of the insertion; position 283- AAACTCTTTATCTCCCCT; 341- AG; 505- TT; 534- TT; and 800 GTA), and / indicate polymorphic sites (in the case of *fil1B* they are putative sites; see text). The only two mutations (sites 201 and 237) found in the coding region are synonymous.

Table 5.4. Summary of the DNA sequences obtained. All the putative loci have been detected in every plant listed, and the individuals from which sequences were obtained are indicated in the table.

Population	Plant	fil1A1 (617 bp)*	fil1A2 (615 bp)*	fil1A3 (615 bp)*	fil1B (618 bp)*	fil1C (636 bp)*
cirrhigerum Ave	2	S	S			
	3	S	S			
	4	S				
	5	S				
	6	S	S			
	7		S			
	8	S				
	9		S			
	10	S				
	11	S				
	12		S			
cirrhigerumGala	1	S				
	2	S	S			
	3	S	S			
	4	S				
	5		S			
	6	S				
	10		S			
	12	S				
	13		S			
	17		S			
cirrhigerumMuel	1	S	S			
	3		S			
	7		S			
	8		S			
	9		S			
	12	S	S			
	13		S			
	14	S				
	15		S			
	16	S				
	19	S				
	21	S				
linkianum	1	S	S		S_G	
	2	S	S		S_G	
					S_T	
	4	S (468 bp)				S
	5	S (468 bp)			S_G	
	6	S (468 bp)			S_G	
	8	S (468 bp)	S (466 bp)		S_G	S
					S_T	
	9					S
	10	S (468 bp)	S (466 bp)			
	11	S	S			
	12	S (468 bp)	S (466 bp)		S_G	S
	14		S (466 bp)		S_G	S
graniticum	1				S_T (436 bp)	
	2	S (468 bp)	S (466 bp)	S (466 bp)		
	4	S (468 bp)	S (466 bp)	S	S_G (436 bp)	
					S_T (436 bp)	
	5				S_T (436 bp)	
	8	S (468 bp)	S (466 bp)		S_T (436 bp)	
digitalis		S (468 bp)	S (466 bp)	S (466 bp)	S_G	S (544 bp)
					S_T	
verbascum		S (468 bp)	S (466 bp)	S (466 bp)		

S indicates individuals for which sequences have been obtained. For *fil1B*, subscripts indicate the nucleotide present at the only variable site (nucleotide position 498; see Results). * The size of the sequenced DNA fragment (primers not included). In some cases, shorter sequences were determined and their size is indicated in brackets. *fil1A1, filA2, fil1A3, fil1B*, and *fil1C* sequences were obtained (after cloning) using primers F1 and F2. The shorter sequences were obtained using the following primer combinations: AGAG and F2 for *fil1A1*; AGAA and F2 (after cloning) for *fil1A2* and *fil1A3*; F1 and BR (after cloning) for *fil1B* ; and CF and F2 for *fil1C* (see Material and Methods).

be a chimera of two different genes. Other aspects of the data suggest that *fil1A*, and *fil1B* sequences may not represent a single gene. I next consider these for each gene in turn.

(i) *fil1A*

Using primers F1 and F2, all of the *A. majus* subsp. *cirrhigerum* individuals yielded *fil1A* sequences (Table 5.4). Shorter sequences were also obtained with specific primers for *fil1A* (see Material and Methods) for seven linkianum and three graniticumB plants. Among all these sequences, only two nucleotide variants were observed (one indel and one nucleotide difference; Fig. 5.1). It was therefore unexpected to obtain three different types of *fil1A* sequences from single individuals. Type 1 has an insertion of two nucleotides (AG), at position 341-342, not present in type 2 and 3; type 2 (like type 1) has a C at position 693 that is a T in type 3 (Fig. 5.1). There must thus be at least two different *fil1A* genes. However, from this result alone it is impossible to determine which types (if any) are allelic.

I next studied two further species: *Digitalis purpurea* and *Verbascum nigrum*. Both yielded a PCR amplification product with the expected *fil1A* size (506 bp) when primers AGAG in combination with F2 were used (these primers should specifically amplify the *fil1A* type 1 sequence; see Materials and Methods). When these amplification products were sequenced and compared with the *Antirrhinum* sequences, they were identical. Similarly, the PCR reaction with AGAA and F2 (for *fil1A2* and *fil1A3*) yielded in both *Digitalis* and *Verbascum* an amplification product of the expected size (504 bp). This was cloned, since this primer combination amplifies both type 2 and 3. However, types 2 and 3 can be distinguished by a fixed difference in a *Tsp*RI

restriction site present in type 3 but not in type 2. Both types of clones were identified for these two species and sequenced. Again the sequences of type 3 clones from *Digitalis* and *Verbascum* were identical to the *Antirrhinum* sequences. Sequences of type 2 from these species were identical to one of two type 2 *Antirrhinum* sequences (Fig. 5.1). The presence of identical sequences in *Antirrhinum*, *Digitalis*, and *Verbascum* for each of these three types strongly suggests that they correspond to three different genes.

(ii) fil1B

fil1B sequences from seven *A. majus* subsp. *linkianum* individuals were obtained with the F1 and F2 primers, and five shorter sequences from four graniticumB individuals were obtained with specific primers for *fil1B* (Table 5.4). In both sets of plants, two types of sequences were found, with several of each being observed (Table 5.4). They differ at a single nucleotide site (position 498, see Fig. 5.1). Both types of sequences were also found in *Digitalis*. There is therefore either an ancestral shared polymorphism or two genes. There are no restriction enzymes that recognise the difference at position 498, and it is difficult to design specific primers that would differentiate these two putative genes.

5.3.2. Variability Within and Between the Antirrhinum Species

For *fil1A1,* all sequences are identical both within and between populations and species (Table 5.4). This includes coding, intron, and 3' flanking non-coding regions. For *fil1A2*, the sequences from cirrhigerumAve, cirrhigerumGala, and linkianum (Table 5.4) were identical in all regions. The only variant found for *fil1A2* is a single variable site in

the 3' flanking non-coding region in two out of the eight sequences from cirrhigerumMuel individuals, but no diversity was found in the coding or the intron regions. For this region π (Nei 1987) is 0.00164 for *A. majus* taken as a whole.

The *fil1C* sequences (from five linkianum individuals and one *Digitalis*) were identical. Because of the extremely low sequence variability found in the studies just described, I did not pursue the study of variability in the other *Antirrhinum* species.

5.4. Discussion

5.4.1. Evidence for a fil1 Gene Family

Data from the *Arabidopsis thaliana* genome sequence has revealed that gene families are common in this plant, despite its small DNA content (Delseny *et al.* 1997; Lin *et al.* 1999; Terryn *et al.* 1999), and many gene families are known in other angiosperms. In the Scrophulariaceae, only a few studies have characterised elements of gene families (Iturriaga *et al.* 1996; Florea and Timko 1997; Delavault *et al.* 1998). Gene families may include both members with very different sequences, and members with similar DNA sequences. Therefore detailed nucleotide studies are needed in order to establish orthology. This is very important in population genetics and evolutionary analyses since both nucleotide polymorphism and divergence can be overestimated if paralogous genes are pooled (Gaut 1998).

In this Chapter, I show that *fil1* (previously described as a single-copy gene; Nacken *et al.* 1991) is a gene family of at least five genes. *fil1A*, *fil1B* and *fil1C* are quite different, and establishing that they are different genes required only PCR reactions in

several individuals of different species using primers specific for these sequences. Digesting the PCR amplification product with several restriction enzymes would also have revealed the presence of the three groups of genes (with all sequences present in every plant examined). This approach can reveal heterogeneity in the PCR product, even for genes more similar than *fil1A*, *fil1B* and *fil1C* genes. However when the different members of a gene family are very similar in sequence (as in the case of *fil1A1*, *fil1A2*, and *fil1A3*) this approach will not work. A substantial amount of molecular work is then required to establish orthology, but the repeated presence of multiple sequences in the same individual, as I found for *fil1A,* is sufficient to establish the presence of more than one gene.

Comparisons between species may also help establish orthology. The rationale is that, after the split of two species, shared neutral polymorphisms are expected to be retained for a limited time before being fixed by genetic drift. Clark (1997) showed that retention for 3.8 *Ne* generations (where *Ne* is the effective population size) is expected only 5% of the time. Such tests require information on both the phylogenetic relationships of the group and the times of origination of the taxa. For the Scrophulariaceae, no fossil data have yet been reported and therefore no direct estimate is available of the time of origin of this family. Divergence times can however, be estimated using molecular data. Molecular phylogenetic studies of several chloroplast genes suggests that *Antirrhinum* and *Digitalis* belong to one clade, while *Verbascum* belongs to a different one (Olmstead and Reeves 1995; Wolfe and dePamphilis 1998; Soltis and Soltis 2000) and this division seems to be one of the oldest within Scrophulariaceae. However bootstrap values supporting these two clades are not high, ranging from less

than 50% (in a combined analysis of two chloroplast genes and the nuclear 18S rDNA genes; Soltis *et al.* 1999) to 74 (in a combined data set of *rbc*L and *ndh*F; Olmstead and Reeves 1995; see Discussion in Chapter 2). Assuming a molecular clock (Zuckerkandl and Pauling 1965) and a divergence of 40 My between Scrophulariaceae and Solanaceae (Xue *et al.* 1996), the age of the split between *Antirrhinum* and *Verbascum* can be estimated from the silent site divergence between orthologous chloroplast genes of *Antirrhinum, Verbascum* and Solanaceae. From the three genes, *ndh*F, *trn*L, and *rbc*L, the average estimated divergence time between *Antirrhinum* and *Verbascum* is 22 My (22.2 My for *ndh*F; 26.2 My for *trn*L; 17.6 My for *rbc*L).

5.4.2. Low Nucleotide Diversity within A. majus and A. graniticum

Out of the five putative loci sequenced, no variability was found for three loci (*fil1A1*, *fil1A3* and *fil1C*) in either the coding, intron or 3' flanking regions, although I sequenced multiple plants sampled from two species, including three different *A. majus* subsp. *cirrhigerum* populations and one population of *A. majus* subsp. *linkianum* (Table 5.4). For the *fil1A2* gene, only one variable site was found in two cirrhigerumMuel sequences in the 3' flanking region. What could account for the low diversity of the *fil1*-like gene family? I consider in turn three possibilities.

(i) *fil1* genes may be located in regions of low recombination

Regional factors, such as recombination rates may influence the observed levels of gene variability. A positive correlation between DNA variation and local recombination rates have been found in several organisms (see Chapter 1). The low level of DNA

variation observed in the *fil1* genes in the *Antirrhinum* species studied could therefore be due to the genes all being located in a region of low recombination. Members of many other plant multi-gene families are clustered in tandem arrays (Yamaguchi-Shinozaki *et al.* 1990; Sutliff *et al.* 1991; Kanazin *et al.* 1996; Rabinowicz *et al.* 1999; Grant *et al.* 2000; Pan *et al.* 2000), so it is possible that the *Antirrhinum fil1* genes all lie in the same genomic region. However, I have no evidence on whether this is true. To test for the effect of local recombination rates on diversity, both genetic and physical maps of the chromosomes are required, and such data are not yet available for any *Antirrhinum* species. However, this explanation seems unlikely since it does not account for the low divergence between different species. Moreover, the same pattern of low diversity was also found in the TCP gene family (Chapter 4).

(ii) Codon usage bias

Non-random usage of synonymous codons is observed in genomes of many species (reviewed by Sharp *et al.* 1993). The patterns of base composition can be roughly divided into two categories: a) trends across the genome at various scales and b) distinctions among functional classes of sites within genes (Akashi *et al.* 1998). Either of these could in principle explain the low variability and low divergence observed in the *fil1* and TCP gene families. A common inverse measure of this bias is the effective number of codons (ENC; Wright 1990). This measure ranges from 20 (only one codon used in each synonymous codon group) to 61 (all synonymous codons equally used in each codon group). The average ENC value for the 52 *Antirrhinum* genes is 53.68 (Table 5.3). The ENC values for the five TCP genes, which show low variability and divergence (Chapter

4) is 56.83, higher than the average. Therefore there is no evidence that these genes are more constrained than the average. Short sequences (less than 100 amino acids), such as the *fil1* genes cannot be analysed because the sample size for each amino acid is small and the results for any amino acid will not be accurate (Moriyama and Powell 1997).

Another simple commonly used estimate of codon usage bias is the G+C level at third position of codons. This index is sensitive neither to rare amino acids nor to short coding regions (Chiapello *et al.* 1998). However it is not very informative since it is a mean value that does not provide details of codon usage in a gene. The GC3 values of the 52 *Antirrhinum* genes examined range from 37.4 to 71.1 % (Fig. 5.2). For the *fil1* GenBank sequence the GC3 content is 56.0 and the *cyc* genes family about 50 (*cyc* 50.5; *cyc1B* 49.2; *cyc2* 49.6; *cyc3* 50.0; *cyc4* 48.7). Therefore neither the *fil1* nor the *cyc*-like genes seem to be highly biased. Furthermore, our observation that variability is low at silent sites, intron sites and in the 3' flanking region, as well as in coding regions, indicates that constraints due to codon usage are not the full explanation of our findings.

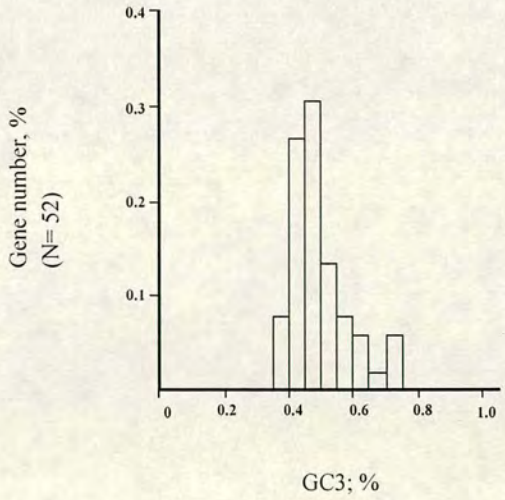


Fig. 5.2. Distribution of 52 coding sequences of *Antirrhinum* according to GC3 levels.

(iii) RNA secondary structure

Sequence conservation in the *fil1* gene family could, in principal, be due to constraints imposed by mRNA structure, which can affect mRNA localisation, stability and translation (Macdonald and Struhl 1988; Mullner and Kuhn 1988; Macdonald 1990; Pandey *et al.* 1994). I therefore tested for the presence of conserved RNA helices in the *fil1* gene sequences (see Material and Methods). I found no evidence for helices.

5.4.3. Low divergence between gene sequences from the different species

Low divergence between species is observed for the five *fil1* genes. The same was found for TCP genes of *Antirrhinum* and *Digitalis* (Chapter 4) and also for four other genes in *Antirrhinum* and *Verbascum nigrum* (see also Chapter 6). It is unclear whether the same factor causes both low polymorphism and low sequence divergence. Location of the genes in regions of low recombination cannot account for the low divergence levels. Genes in regions of low recombination tend to have low bias (Kliman and Hey 1993; Comeron *et al.* 1999; Duret and Mouchiroud 1999) and should therefore show more synonymous substitutions per site than genes in regions of normal recombination (Munte *et al.* 1997). It is also unlikely that codon bias could explain both low polymorphism and divergence since there is no sign of different rates of divergence in regions of putatively different constraint (such as 5' flanking regions, intron, synonymous and replacement sites).

Concerted evolution also cannot account for the observed low divergence. The substitution rate of alleles in a multigene family is expected to be the same as for single-copy genes (Nagylaki 1983; Walsh 1985). Nor does it account for low polymorphism

within species. Gene conversion and/or unequal crossing-over, particularly if the genes are a tandem array, could explain the similarity of the different *fil1*-like genes. Theoretical models show that these mechanisms can be important in maintaining sequence homogeneity among repeated genes (Nagylaki and Petes 1982; Ohta 1981, 1984) which may be almost identical within a species (see, for instance, Zimmer *et al.* 1980). A detailed characterisation of the number and location of the repeated genes would be of interest, but such data are not available for any of the genes under study, for either *Antirrhinum* or any other gene family in the Scrophulariaceae. However, there is evidence for inter-locus exchange in the alcohol dehydrogenase gene family in grasses (Gaut *et al.* 1999) and also in the chalcone synthase genes of millet (Oberholzer *et al.* 2000) and it may therefore be an important aspect of plant genome evolution.

A possible resolution of the puzzle presented by these data is that these genes or species have an unusually low mutation rate. Since the recombination rate in the region where a gene is located does not affect the silent substitution rate, each gene gives an independent estimate of the mutation rate, assuming neutrality (Aquadro *et al.* 1994). It is not possible to get absolute rate estimates, because there is no good estimate of the age of the Scrophulariaceae. Nevertheless, since the *Antirrhinum* and *Verbascum fil1*-like genes are very similar, even a recent origin of the Scrophulariaceae would give rates much lower than the lowest estimates yet published for other nuclear genes (in monocotyledonous palms, Wolfe *et al.* 1989; Gaut *et al.* 1996 and the dicotyledonous plant *Gossypium*, Small *et al.* 1999).

Chapter 6

Low rates of silent substitution in nuclear genes of two distantly related Scrophulariaceae (*Antirrhinum* and *Verbascum*)

6.1. Introduction

In *Antirrhinum,* detailed studies of DNA sequence diversity and divergence of genes of TCP and *fil1* gene families have revealed a puzzling pattern of little variation within species, and low divergence between several *Antirrhinum* species and *Digitalis purpurea* (Chapter 4 and 5). For two *fil1* genes where orthologous *Verbascum nigrum* sequences were obtained, low levels of divergence were also observed (Chapter 5). This was unexpected, though low levels of sequence divergence have sometimes been found in other morphologically distinct plants (Hodges and Arnold 1994, 1995; Aceto *et al.* 1999; Wagstaff *et al.* 1999). The species studied differ in morphology and life forms, and the *Antirrhinum* species studied have different breeding systems. Moreover, allozyme divergence has been found in *A. lopesianum, A. mollissimum* and *A. microphyllum*

(Mateu-Andres 1999) and other Scrophulariaceae (Elisens and Crawford 1988; Hamrick and Godt 1990; Ritland 1989; Elisens 1992; Elisens and Nelson 1993).

In the case of the intronless *cyc*- like genes, only the coding region was analysed, and it was therefore possible that the low diversity and divergence observed was due to an unusually high level of constraint on the coding sequences. However in the case of the *fil1*- like genes, similar results were obtained for both intron and 3' non-coding regions. This suggests that an unusual level of constraint is not the explanation for the low diversity and divergence. Furthermore, both *fil1* and *cyc*- like genes have low codon usage bias, compared with other *Anthirrinum* genes (Chapter 5), so there is no evidence suggesting severe constraints in their sequences.

The low divergence observed for these two gene families between the sequences in *Antirrhinum*, *Digitalis* and *Verbascum* could be partially explained if the species analysed were closely related. In the glasshouse, species of *Antirrhinum (A. meonanthum, A. majus* subsps *linkianum, A. majus, A. molle, A.glutinosum, A. latifolium, A. hispanicum, A. barrelieri*) can hybridise easily, yielding viable progeny (Mather 1947; Harrison and Darby 1995; Rothmaler 1956) and this is common in the Scrophulariaceae (Elisens 1992). Phylogenetic studies using chloroplast gene sequences are consistent with *Digitalis* and *Antirrhinum* being closely related (both belong to the Scroph II clade). However, *Verbascum* (which belongs to the Scroph I clade) should not be closely related to either *Antirrinum* or *Digitalis* (Olmstead and Reeves 1995; Wolfe and dePamphilis 1998; Soltis and Soltis 2000; but see Soltis *et al.* 1999 and Discussion in Chapter 2).

It is, therefore, unclear why the *fil1* and *cyc* gene families show low divergence between *Antirrhinum* and *Verbascum.* It is also unknown whether this is general, or a feature of a particular group of genes.

In order to study the generality of the low divergence among nuclear genes of the Scrophulariaceae, I searched GenBank for pairs of homologous genes from *Antirrhinum* and at least one other in species of Scrophulariaceae in orther to see whether they could be orthologous. Pairs were found from *A. majus*, *Digitalis lanata*, *Striga hermonthica*, *Craterostigma plantagineum*, *Paulownia kawakamii*, *Torenia hybrida, Alonsoa meridionalis* and *Asarina barclaiana* species. None of the seven gene pairs found could be considered orthologous, suggesting that gene families may be common. I also obtained further sequences in *Antirrhinum* and *Verbascum* for the *fil2*, *far*, *globosa* and *Adh* genes. *fil2* is a flower-specific gene that encodes a protein of the extracellular matrix with a 10 amino acid LRR signal peptide at the N-terminus (Steinmayr *et al.* 1994). BlastX search reveals that *fil2* is a polygalacturonase gene. Farinelli (*far*) and *globosa* are floral homeotic genes that control petal and stamen development. *far* belongs to class C and *globosa* to class B of MADS-box genes (Tröbner *et al.* 1992; Davies *et al.* 1999). Here, I show that in *Antirrhinum* all these genes belong to gene families. For four genes belonging to the *fil2*, *far*, and *globosa* gene families, identical sequences were obtained from *Verbascum*. I also sequenced alcohol dehydrogenase (*Adh*) genes in these taxa, and again found a gene family. For the *Adh* genes, an apparent history of repeated gene duplication and loss of elements made the establishment of orthology impossible.

6.2. Material and Methods

Plant material and DNA extraction

Leaves of *Antirrhinum majus* ssp *cirrhigerum* were collected in the field in the North of Portugal (Aveiro) in 1997 (Table 3.1 in Chapter 3). *Verbascum nigrum*, *V. thapsus* and *Digitalis purpurea* leaves were collected from wild individuals growing on the Edinburgh University campus. Genomic DNA was prepared from leaves of individual plants using the method of Ingram *et al.* (1997).

PCR amplification

Based on the GenBank sequences of *fil2, farinelli (far), and globosa* (accession numbers X76995, AJ239057, and X68831 respectively), primers were designed for these genes (Table 6.1). The regions analysed for each of these genes are shown in part A of Figures 6.1- 6.3. *fil2* is known to be a gene family in a number of species including tomato, melon, maize and willow (Hadfield and Bennett 1998; Futamura *et al.* 2000). Homologues of *far* and *globosa* have also been described as belonging to gene families (Kramer *et al.* 1998; Yu *et al.* 1999; Theissen *et al.* 2000). However, among *Antirrhinum* sequences in GenBank, only *far* shares homology with any other *Antirrhinum* gene (in this case *plena*). Therefore primers were designed to amplify only *far.* Since our work revealed that these genes are members of gene families in *Antirrhinum*, additional primers were designed based on the new sequences obtained (see Results). To avoid amplifying multiple members of these gene families, semi-nested PCR (Cubas *et al.* 1999b) was carried out on the product of the initial PCR reactions with further internal primers.

Antirrhinum Adh sequences were not available in GenBank. Therefore, conserved regions of 20 bp were identified based on the alignment of *Adh1*, *Adh2*, and *Adh3* genes from other dicotyledon species (*Solanum tuberosum* M25154, M25153, M25152; *Leavenworthia stylosa* AF037564; AF037558; AF037560; *Leavenworthia crassa* AF037563). Primers adA and adB (see Table 6.1) were designed for these regions. These primers amplify a small region (392 bp) corresponding to part of exon 4 in *Arabidopsis thaliana* (D63464).

Table 6.1. Primers used in this study

Gene	Primers		Position relative to the start codon
fil2	Fil2A	5' CGACTTCCACCTCTTCTCC 3'	216
(X76995)	Fil2B	5' TCCAACGACGATAAACTGT 3'	1042
	fil2-1R	5' GGAGGGATCGAGCCACTAAA 3'	
	fil2-2R	5' GGAGGGATCGAGCCACTGAG 3'	
far	FarF	5'ACAAGAAAGCATCATCAGA 3'	3218
(AJ239057)	FarR	5' CTTGGCTCGCAGATACTGG 3'	4116
	Far1 F	5' CTCAGGTACTTTAGTTTTT 3'	
globosa	GloF1	5' ATTTTGTTTTGAGGGACTA 3'	1434
(X68831)	GlobR	5' GGCTTTGATTCTCCTCCTC 3'	2279
	Glogape	5' CTTTCCCATCATTTTTATTA 3'	
Adh	adA	5' ATTGTGGAGAGTGTTGGAGA3'	
	adB	5' CCAAGTCCAAAAACAGCAAC 3'	
	v1	5' AAAGGAAGGAGACACAGTT 3'	

Standard amplification conditions were 35 cycles of denaturation at 94° C for 30s, primer annealing at 48° C for 30 s, and primer extension at 72° C for 2 min. Because I was working with very similar sequences, it was important to be extremely careful to avoid contamination. Stringent controls were therefore continually included in the work. Negative controls (PCR cocktail with no genomic DNA added) were included in all PCR amplifications and never yielded PCR products. Also the PCR amplification was repeated with at least six individuals of *A. majus* ssp *cirrhigerum*, and two of them were always sequenced. Since only one *Verbascum nigrum* individual was used, DNA from several leaves was extracted independently, and treated as different samples. The PCR of these samples always gave the same amplification products.

We routinely checked the PCR amplification products for homogeneity by digestion with several four-cutter restriction enzymes. If the number and/or the size of the bands obtained after digestion is not compatible with that of the reference sequence (from which the primers were designed), the amplification product was classified as heterogeneous. In such cases, I cloned the products and screened several colonies until several of each of the types previously revealed were found, and determined their DNA sequences. Cloning was performed using the TA cloning kit (Invitrogen). If more than one band was systematically obtained, I always cloned and sequenced all of them. Because differences can arise from nucleotide mis-incorporation during amplification, I determined the DNA sequence of at least three different colonies and obtained a consensus sequence. DNA sequencing was performed with an Applied Biosystems model 377 DNA sequencing system with the ABI PRISM BigDye cycle-sequencing Kit

(Perkin Elmer), using specific primers or the primers for the M13 forward and M13 reverse priming sites of the pCR 2.1 vector.

Analyses of the sequences

The DNA sequences were deposited in GenBank (accession numbers AF307068-307071 for *fil2-1*,AF307072-307075 for *fil2-2*, AF307063-307064 for *farL*, AF307065-307067 for *farS*, Af307076-307078 for *globosa1*, and AF307054-307062 for *Adh*). The nucleotide sequences to be compared were aligned using ClustalX v. 1.64b (Thompson *et al.* 1997), and minor manual adjustments were performed using SeqPup v. 0.6f. The numbers of synonymous and nonsynonynous differences between pairs of sequences were calculated using the DnaSP software (Rozas and Rozas 1997). Intron/exon boundaries for the genes analysed were deduced by comparison with the homologous genes deposited in GenBank. Divergence estimates were corrected for multiple hits using Jukes-Cantor correction (Jukes and Cantor 1969). Neighbor-joining trees were generated for the *Adh* genes with MEGA version 1.01 (Kumar *et al.* 1994).

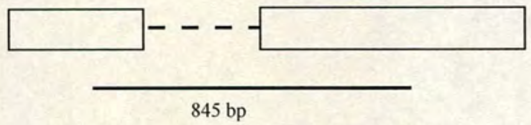
To find homologous genes of Scrophulariaceae, I used BlastX searches, which use only protein coding regions (Altschul *et al.* 1997). Out of 62 sequences available from species other than *Antirrhinum*, six had significant amino acid similarity with *Antirrhinum* sequences, and one with another non-*Antirrhinum* gene (Table 6.2). For comparative purposes, I also included a sequence from a species in the family of Geraniaceae, Solanaceae or Fabaceae.

6.3. Results

6.3.1. Evidence that fil2, far, globosa and Adh are members of gene families

i) *fil2*

Primers Fil2A and Fil2B (Table 6.1) amplify a PCR product of the predicted size (845 bp) from two *A. majus* subsp *cirrhigerum* plants (Fig. 6.1). Digesting this band with several restriction enzymes revealed that it was a mixture of two types of DNA sequences. I therefore cloned this PCR product and sequenced several clones corresponding to both types of DNA sequences from both individuals. The sequences fell into two different types. Each individual contributed one sequence of each type, but the sequences from these two individuals were identical within each type. Blast searches with these two types of sequences revealed high nucleotide sequence similarity (94% and 95% nucleotide identity, respectively) with the *fil2* gene (accession number X76995). Between the two types of DNA sequences there are 17 nucleotide differences (nine nonsynonymous and eight synonymous) and one four bp indel in the putative intron region (Fig. 6.1). Both types of sequences were amplified from all of a sample of 12 *A. majus* ssp *cirrghigerum* individuals analysed, strongly suggesting that they represent two different genes. I therefore named the sequences *fil2-1* and *fil2-2*. The *fil2* sequence from GenBank (accession number X76995) has several differences relative to both *fil2-1* and *fil2-2*. For *fil2-1* there are 11 nonsynonymous, 11 synonymous, and five nucleotide differences in the putative intron, plus eight indels, and for *fil2-2* four nonsynonymous, five synonymous, and five nucleotide differences in the putative intron, plus seven indels (Fig. 6.1).



A

Exon1

Sequence	5	28	84	99	107
fil2-1					
cirrhigerumAve1 cirrhigerumAve3	A	G	C	A	T
Verbascum nigrum *V. thapsus*	.	.	.	.	.
fil2-2					
cirrhigerumAve1 cirrhigerumAve3 *Verbascum nigrum* *V. thapsus*	.	.	.	.	A
GenBank sequence *A. majus*	G	A	T	G	T

Intron 1

Sequence	116	158	178	189	195	228	236	243	250	252	258	269	296	304	366
fil2-1															
cirrhigerumAve1 cirrhigerumAve3	T	A	G	A	A	-	-	T	G	A	G	T	-	A	*i*(4)
Verbascum nigrum *V. thapsus*	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
fil2-2															
cirrhigerumAve1 cirrhigerumAve3 *Verbascum nigrum* *V. thapsus*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-
GenBank sequence *A. majus*	-	G	A	-	-	*i*(2[a])	*i*(2[b])	A	-	G	T	.	*i*(7)	.	.

Exon 2

Sequence	384	396	426	429	439	451	462	470	478	483	493	495	520	542	609	660	663	678	683	684	732	754
fil2-1																						
cirrhigerumAve1 cirrhigerumAve3	C	A	C	T	T	G	C	A	T	G	T	T	G	G	T	A	T	T	T	A	C	A
Verbascum nigrum *V. thapsus*	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
fil2-2																						
cirrhigerumAve1 cirrhigerumAve3 *Verbascum nigrum* *V. thapsus*	.	G	G	C	G	.	.	G	C	A	C	C	.	A	.	T	A	A	C	G	.	T
GenBank sequence *A. majus*	T	G	.	C	G	.	.	G	C	A	C	C	A	A	.	T	A	A	C	G	T	T

B

Fig. 6.1. (A) Schematic diagram of *fil2* GenBank sequence (*A. majus*). Boxes represent exons and the dotted line is the single intron (modified from Steinmayr *et al*., 1993). The 845 bp region analysed is indicated by a line. (B)Variable sites of the *fil2*-like sequences. Dots represent the same nucleotide as in the first sequence, dashes indicate deletions, *i* indicate insertions (the numbers in brackets represent the sizes of the insertions; position 228- GC; 236- GG; 296- ATATTAA; and 366-TAAG. Sequence names are according to the population code for *Antirrhinum majus* subsp. *cirrhigerum* and numbers after the code name are sample codes (Chapter 3).

Based on two fixed differences between *fil2-1* and *fil2-2* at positions 493 and 495 (Fig. 6.1), specific primers were designed for these genes (Table 6.1). Using these primers, PCR amplification products were also obtained from genomic DNA of *Verbascum nigrum* and *V. thapsus,* and their sequences determined. For both *fil2-1* and *fil2-2*, the *Verbascum nigrum* and *V. thapsus* sequences are identical. When the *Antirrhinum* and *Verbascum* sequences (481 bp long) were compared, only one difference was found in *fil2-1,* and none in the *fil2-2* gene (Fig. 6.1). The presence of these sequences in *Verbascum* also supports the conclusion that there are at least two *fil2* genes.

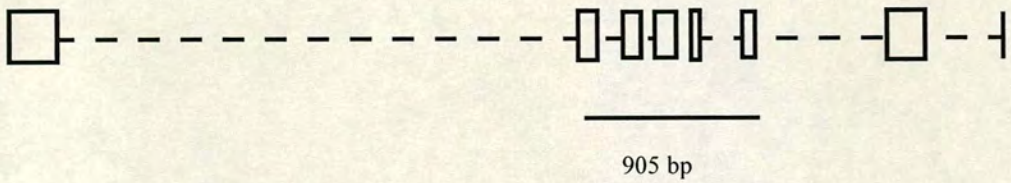
ii) *far*

Two bands of different sizes (905 bp, called L and 731 bp, called S) were amplified from *A. majus* subsp. *cirrigherum* using primers FarF and FarR (Table 6.1). Both bands were always obtained in six individuals from this population. For two individuals both L and S were cloned and sequenced in order to establish whether these bands were specific amplification products. The sequences from these two individuals were identical within each band. The analysis of these sequences revealed that both L and S sequences share significant nucleotide similarity (96% and 88% nucleotide identity, respectively) with the GenBank sequence *far* (accession number AJ239057). Comparison of the L and S sequences, putative coding regions indicated 15 differences at silent sites, and seven at replacement sites (Fig. 6.2). In the putative introns, differences in size make the alignment of L and S sequences ambiguous (the length of intron one of the L sequences is 171, versus121 bp in the S sequences; the intron three lengths are 80 versus 97 bp, and

the intron four lengths are 254 versus 106 bp for L and S, respectively). Based on the S sequence, a new primer spanning the end of exon two and the beginning of intron three was designed to specifically amplify this sequence, and PCR amplification was done using *Verbascum nigrum* DNA (Table 6.1). The sequence of the *Verbascum* S band (662 bp long) is identical to that of *Antirrhinum*. The number of differences between the sequences of the two *Antirrhinum* types with different band sizes, coupled with the similarity of the *Antirrhinum* and *Verbascum* S sequences, strongly suggests that the L and S types of sequences are two different genes (*far-S* and *far-L* genes). The *Antirrhinum majus far* GenBank sequence is most similar to the *far-L* gene, but there are several differences between them (5 synonymous nucleotide differences; 18 intron nucleotide differences and 5 intronic indels), and it is possible that they are not allelic (Fig. 6.2).

iii) *globosa*

Primers GloF1 and GloR (Table 6.1) amplified a PCR product of 994 bp in *Antirrhinum majus* subsp *cirrhigerum*. This amplification product was not of the expected size (994 bp, rather than the expected 864 bp long product based on the sequence in GenBank), but it was observed in all six individuals studied in this species. There is evidence for both ancient and recent duplications of *globosa*-like genes in other angiosperms (Kramer *et al.* 1998), though *globosa* has not been described as a member of a gene family in *Antirrhinum*. I cloned and sequenced this product from two individuals and both



A

Exon	2				3							4												5			
	38	39	50	53	234	252	256	261	275	276	288	392	394	397	411	415	416	427	444	445	451	456	463	473	576	580	591	603
far-L																												
cirrhigerumAve3 cirrhigerumAve13	T	A	A	C	T	G	C	T	C	T	G	T	G	A	C	G	T	A	G	C	C	G	T	A	G	G	G	G
far-S																												
cirrhigerumAve3 cirrhigerumAve13 *Verbascum nigrum*	A	G	C	T	C	.	T	C	T	.	A	C	T	T	G	A	A	T	C	T	T	A	C	C	A	A	A	A
GenBank sequence *A. majus*	.	.	.	.	.	T	.	.	T	C	.	.	.	.	.	.	.	G	.	.	T	.	.	.	.	.	.	.

B

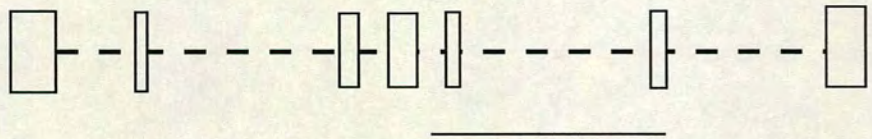
Fig. 6.2. (A) Schematic diagram of the *far* GenBank sequence (*A. majus*). Boxes represent exons and the dotted lines introns (modified from Davies *et al.*, 1999). The 905 bp region analysed is indicated by a line. (B)Variable sites in the putative coding region of *far*-like sequences. Definitions as in Fig. 6.1.

sequences were identical. Analysis of these sequences revealed that they share significant nucleotide similarity in both the intron and coding region with the GenBank *globosa* sequence (accession number X68831; > 95% nucleotide identity for both regions). The difference in size is mainly due to a duplication of a 129 bp intron region of the GenBank sequence. Between our sequence and the one deposited in GenBank, there are 17 nucleotide differences (two of them synonymous, and 15 in the intron; Fig. 6.3) plus five indels, out of the 826 sites compared.

Once again, gene duplication would be indicated if a distantly related species has a sequence similar to the one that I obtained, and divergent from the one in GenBank. To test for the presence of a similar sequence in *Verbascum nigrum*, I designed a new primer (Glogap; see Table 6.1) for the region of the gene that is duplicated relative to the GenBank sequence (see above; Fig. 6.3). Since this region is represented twice in the target sequence, two bands with different sizes (512 bp and 641 bp) are expected, and both were obtained both in *Verbascum* and *Antirrhinum* using Glogap and GloR primers. In the 622 bp region analysed, the *Verbascum* and *Antirrhinum* sequences were identical.

iiii) *Adh*

Using primers adA and adB (Table 6.1) a PCR product with the expected size (392 bp) was obtained from both *A. majus* subsp *cirrghigerum* and *Verbascum thapsus*. These primers were designed for sequences conserved in three *Adh* genes of two very distantly related genera (see Material and Methods). Therefore, this PCR product is expected to be heterogeneous. The amplification product was cloned from both species. Several



A

	Intron 4			Exon 5		Intron 5																				
	31	34	37	97	100	148	153	155	170		189	216	276	281	294		315		346	350		393	456		458	629	639
GenBank sequence *A. majus*	A	T	G	T	C	A	A	T	*i*(3)		T	T	A	C	-		-		A	*i*(4)		C	-		C	C	G
globosa1																											
cirrhigerumAve1 cirrhigerumAve3 *Verbascum nigrum*	C	A	T	C	T	T	T	C	-		A	A	T	A	*i*(1)		*i*(8)		G	-		T	*i*(129)		T	T	G

B

Fig. 6.3. (A) Schematic diagram of the *globosa* GenBank sequence (*A. majus*). Boxes represent exons and the dotted lines introns (according to Tröbner *et al.*, 1992). The 864 bp region analysed is indicated by a line. (B) Variable sites of *globosa*-like sequences. Definitions as in Fig. 6.1. *i* indicates insertions (position 170- ATA; 294- A; 315- CTAACACA; 350- TTAA; and 459- 129 bp long duplication that is identical to the 323- 455 region except for a three bp deletion).

clones (on average 30) for each species were digested with various restriction enzymes, namely *Aci*I; *Alu*I, *Rsa*I, and *Dde*I. In *Antirrhinum,* this approach revealed three types of clones, and in *Verbascum* two types. For both species, several clones of each type were sequenced.

In *Antirrhinum*, five different sequences were obtained from the same individual. Blast search revealed that all sequences share the highest amino acid similarity (>62 % amino acid identity) with an *Adh*-like sequence from *S. tuberosum* (accession number X92179). Since *Antirrhinum* is a diploid species, the presence of five different sequences implies the presence of at least three genes.

An *Adh* gene tree was obtained using the sequences obtained here and those previously deposited in GenBank for Solanaceae (Fig. 6.4; note that no *Adh* sequences from species in the Scrophulariaceae are available in GenBank). The sequences of the three types of clones are in different branches, which are supported by high bootstrap values, compatible with our interpretation that they are different genes. I called them *Adhant1*, *Adhant2*, and *Adhant3*. It should be noted the nomenclature used for the *Adh* gene family does not imply orthology with other *Adh* plant genes with the same numbers. The pairwise mean difference per synonymous sites varies from 0.192 (comparing *Adhant1* and *Adhant2*) to 1.41 (*Adhant2* versus *Adhant3*). The two *Adhant1* sequences from the single individual studied are similar and could be allelic (the mean number of synonymous substitutions per synonymous site is 0.0247). The same applies to the two *Adhant2* sequences (the number of synonymous substitutions per synonymous site is 0.0244). These values are much higher than the average for TCP and *fil1* genes of *Antirrhinum* species (Chapter 4 and 5).

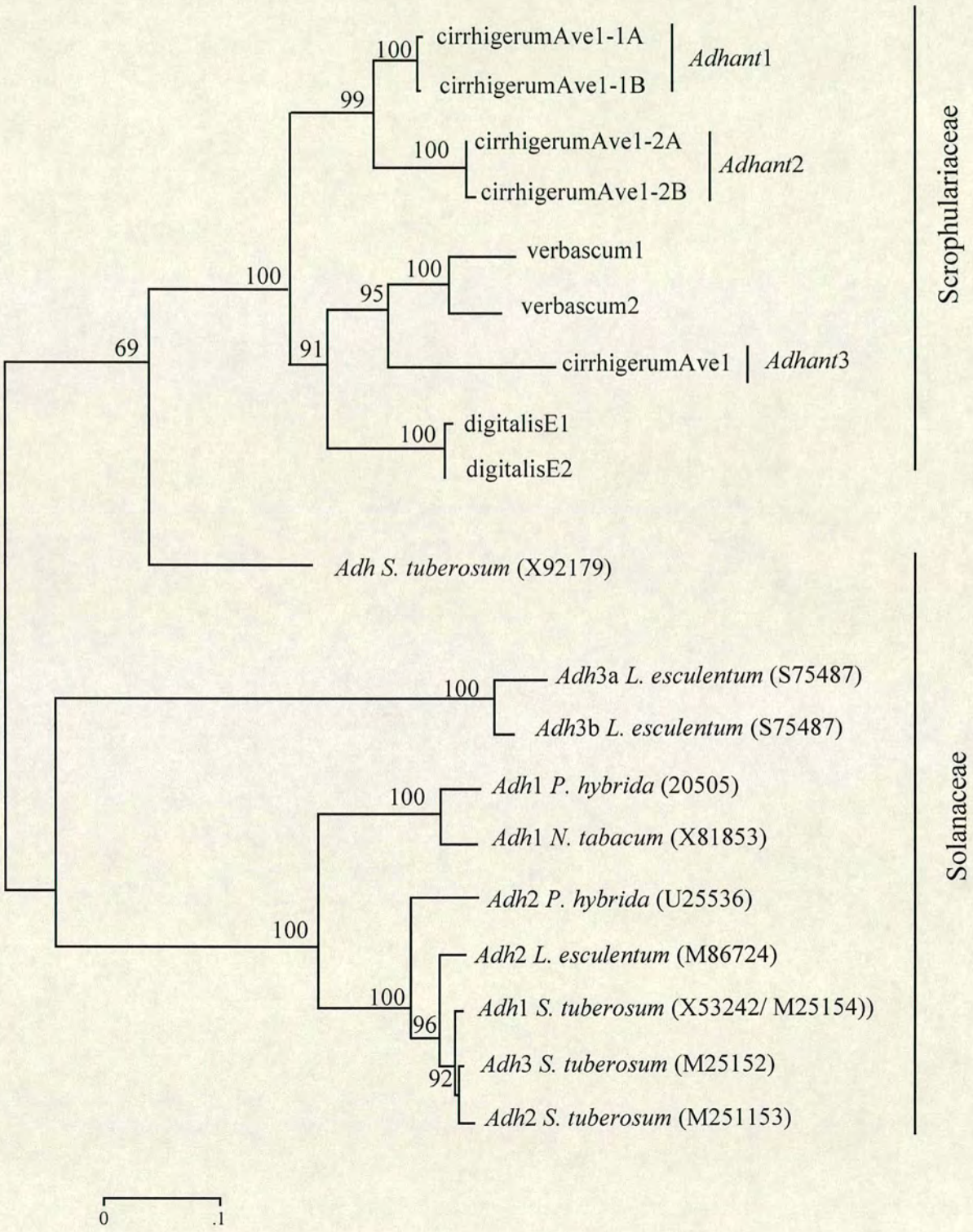


Fig. 6.4. Unrooted neighbor-joining tree using the Kimura two-parameter distance, showing the relationships among the *Adh* genes of Scrophulariaceae and Solanaceae. Bootstrap replicates supporting the branches are shown for values greater than 68%.

In *Verbascum*, the RFLP survey revealed only two types of clones. For one individual, consensus sequences for each type were obtained. These appear to represent two different genes that arose from an ancient duplication, since they differ by an average of 0.453 synonymous substitutions per synonymous site. The difference in copy number in *Antirrhinum* and *Verbascum* makes it difficult to establish which (if any) of these genes are orthologous (Fig. 6.4).

It is unlikely that other *Adh* genes exist in the *Antirrhinum* genome that are more closely related to the *Verbascum* sequences than the ones amplified with the primers adA and adB. Using another primer pair (v1 and adB, in which v1 was designed based on the verbascum1 sequence), no amplification product was obtained in *Antirrhinum*, although a range of different annealing temperatures, down to 45°C was tested. A PCR product of the expected size was always obtained in *Verbascum*. Therefore *Adh* copy numbers seem to have changed over evolutionary time. Additional evidence comes from *Digitalis*. Based on chloroplast phylogenies, this species seems to be more closely related to *Antirrhinum* than to *Verbascum* (Olmstead and Reeves 1995; Wolfe and dePamphilis 1998; Soltis and Soltis 2000; but see Discussion in Chapter 2).When ten *Digitalis* clones (obtained with primers adA and adB from one individual) were restricted with four restriction enzymes, only one type of clone was found. When these were sequenced, two different sequences (digitalis1 and digitalis2; Fig. 6.4) were revealed. They differ at four nucleotide positions (two of them nonsynonymous; the mean number of synonymous substitutions per synonymous site is 0.0243), and they may be allelic.

These *Digitalis* sequences do not cluster with any of the *Antirrhinum Adh* sequences (Fig. 6.4).

6.4. Discussion

6.4.1. The Antirrhinum fil2, far, globosa and Adh genes are members of gene families

Here, I show that in *Antirrhinum, fil2*, *far*, *globosa* and *Adh* genes are members of gene families. For *fil2*, *far*, and *globosa,* the genes described here are probably not allelic with the ones from *Antirrhinum majus* that are in GenBank (see Results). It is unclear why the primers used here did not amplify sequences similar to those in GenBank, since they were designed based on these sequences. However, the total copy number for these gene families is unknown in both *Antirrhinum* and *Verbascum* and could be large. It is also unclear why no evidence for a gene family was found when these genes were first described (Tröbner *et al.* 1992; Steinmayr *et al.* 1994; Davies *et al.* 1999) given that cDNA and genomic clones were sequenced.

6.4.2. Copy number fluctuation in the Adh gene family

The unrooted neighbor-joining tree using the Kimura two-parameter distance, shows the relationships among the *Adh* genes of Scrophulariaceae and Solanaceae (Fig. 6.4). Judging from divergence between pairs of sequences, some gene duplication events are ancient, which makes it difficult to find an appropriate outgroup to root the tree. A complex history of gene duplication and loss makes it difficult to infer orthology among

the *Adh* genes of *Antirrhinum*, *Verbascum*, *Digitalis*, and the Solanaceae. Since the genes from the Solanaceae and Scrophulariaceae do not cluster together, a number of gene duplications seem to have occurred after the split of these two families. For any given gene, orthologous loci should reflect the relationships between the species being compared. Therefore this gene tree implies a minimum of three independent duplications, one in each lineage of the Scrophulariaceae analysed here. The same reasoning can be applied to the Solanaceae. In the Solanaceae, a minimum of five duplications seem to have occurred. Although the number of *Adh* genes is generally low among angiosperms (usually two or three loci) as in our data, phylogenetic analyses has revealed similar repeated gains and losses of genes in other genera as well, even among closely related species (Gaut *et al.* 1996, 1999; Morton *et al.* 1996; Clegg *et al.* 1997; Small and Wendel 2000). It should be noted that I have not characterised the *fil2*, *far*, and globosa gene families in detail in either *Antirrhinum* or *Verbascum*. Changes in copy number would therefore not be detected. However, unlike the *Adh* gene family, each of the other genes studied yielded similar sequences from *A. majus* subsp. *cirrhigerum* and *Verbascum*. It thus seems clear that different loci exhibit different evolutionary patterns.

Although more data are available for *Adh* than for any other plant nuclear gene family, gene duplications and losses have also been described in other gene families, for instance *chs* (an enzyme of the chalcone synthase pathway; Clegg *et al.* 1997; Durbin *et al.* 2000; Oberholzer *et al.* 2000), *rbc*S (Clegg *et al.* 1997), *glo*-like and *def*-like (MADS-box genes; Kramer *et al.* 1998; Theissen *et al.* 2000) and nucleotide binding site-leucine rich repeat resistance genes (R-genes; Meyers *et al.* 1999; Pan *et al.* 2000). The frequency with which such duplications occur is difficult to estimate. Few studies

describe the total copy number for such gene families in a group of species. Based on the average number of independent lineages inferred for the *Adh*, *Chs* and *rbc*S gene families, within Poaceae, Asteraceae, Fabaceae and Solanaceae, Clegg *et al.* (1997) suggest a faster rate of duplication for the *Chs* and *rbc*S gene families than for the *Adh* gene family. These authors also suggested that the appearance of new gene copies is infrequent within families. However this may be the result of poor species representation in each family. For instance, for *Adh* only one genus was included for each of the families Malvaceae, Vitaceae, Asteraceae, and Pinaceae. Our data suggest duplications of *Adh* within the Scrophulariaceae, similar to those inferred in the genus *Gossypium* (Small and Wendel 2000). If copy number fluctuation is frequent in gene families, inferences about orthology may be difficult even among closely related species. For instance, I could not establish orthology between the *Adh* genes of *Antirrhinum* and *Verbascum*. Divergence between the sequences of these species cannot therefore be estimated.

6.4.3. Orthology among Scrophulariaceae GenBank sequences

In order to get additional evidence on sequence divergence between species of Scrophulariaceae, I also searched GenBank for pairs of nuclear genes that are homologous. For any given gene, the K_s (and perhaps K_a) values for orthologous loci should reflect the relationships between the species being compared, and more distantly related species should have larger K_s and K_a values than more closely related species. Seven genes could be examined (see Material and Methods). By this criterion, the data in Table 6.2 clearly indicate that four of the genes (*ACS1*, *PHYA*, *GAPDH*, and the MADS-box transcription factor) compared cannot be orthologues; K_s values between species

within the Scrophulariaceae are as high or higher than between Scrophulariaceae and members of other plant families. For three genes (*Chs*, *TFNS5* and *SUT1*) orthology is not excluded. However, for these genes, substitution at silent sites is saturated (on average every synonymous site has been substituted at least once) in the species compared, and K_s values are as high as for two of the first four genes. It is therefore very unlikely that any of these pairs are orthologous genes. No phylogenetic information is available for *Asarina* (Olmstead and Reeves 1995; Wolfe and dePamphilis 1998; Soltis and Soltis 2000), but *Antirrhinum*, *Digitalis*, and *Alonsoa* have been assigned to the same clade (Scroph II), and are therefore not expected to have highly diverged sequences of orthologous genes.

6.4.4. Hypotheses that could account for the low divergence observed for fil2, far, and globosa

Concerted evolution by gene conversion may retard divergence among paralogous sequences within a genome. However this should not affect the level of divergence of orthologous genes (Ohta 1981; Nagylaki and Petes 1982; Arnheim 1983; Ohta 1984). A recent duplication could also explain the similarities among genes within species, but cannot account for low divergence of homologous genes between distantly related species (because the time since the split of such species implies that sequence differences should have accumulated). Therefore gene conversion or a recent duplication could in principle explain the observation of very similar genes within species, but fails to explain the low divergence between species of Scrophulariaceae.

Table 6.2. Comparison between GenBank sequences of nuclear genes from *Antirrhinum,* other Scrophulariaceae, and more distantly related plant species.

Gene		region analysed (bp)	Scrophulariaceae	Outgroup	Evidence of gene families in other plant species
ACS1	*Antirrhinum majus* (AF083814)	317	*Striga hermonthica* (AF090351) K_s 2.27 K_a 0.23	*Pelargonium hortorum* (Geraniaceae) (U17231) K_s 2.38 K_a 0.18	Wang and Arteca 1995
PHYA	*Antirrhinum majus* (U08142)	332	*Digitalis lanata* (AJ002525) K_s 1.45 K_a 0.05	*Sophora affinis* (Fabaceae) (U78835) K_s 1.10 K_a 0.07	Lavin *et al.* 1998; Alba *et al.* 2000
GAPDH	*Antirrhinum majus* (X59517)	1114	*Craterostigma plantagineum* (X78307) K_s 1.11 K_a 0.08	*Petunia* (Solanaceae) (X60346) K_s 1.04 K_a 0.07	Gianfagna *et al.* 1998
MADS-box transcription factor	*Antirrhinum majus* (Y10750)	579	*Paulownia kawakamii* (AF060880) K_s >4 K_a 0.56	*Solanum tuberosum* (Solanaceae) (AF008651) K_s 2.51 K_a 0.57	Shinozuka *et al.* 1999

Table 6.2. Continued

Gene		region analysed (bp)	Scrophulariaceae	Outgroup	Evidence of gene families in other plant species
Chs	*Antirrhinum majus* (X03710)	1165	*Digitalis lanata* (AJ002526) K_s 1.44 K_a 0.04	*Perilla frutescens* (Lamiaceae) (AB002582) K_s 3.45 K_a 0.06	Gong *et al.* 1997; Fukada-Tanaka *et al.* 1997; Durbin *et al.* 2000
TFNS5	*Antirrhinum majus* (AB028151)	1566	*Torenia hybrida* (AB028152) K_s 1.55 K_a 0.15	*Glycine max* (Fabaceae) (D83968) K_s 3.19 K_a 0.46	Schopfer *et al.* 1998
SUT1	*Alonsoa meridionalis* (AF191025)	1467	*Asarina barclaiana* (AF191024) K_s 1.06 K_a 0.14	*Nicotiana tabacum* (Solanaceae) (X82276) K_s 1.78 K_a 0.21	

Either a recent evolution of the Scrophulariaceae or else a low substitution rate could account for the low divergence observed. I now consider whether there is evidence to support either of these hypotheses.

(i) Recent origin of the *Antirrhinum* and *Verbascum* species

The age of the split between *Antirrhinum* and *Verbascum* is unclear, since no fossils are available for Scrophulariaceae. Based on the ribulose bisphosphate carboxylase large subunit (*rbc*L) gene sequences of Solanaceae and Scrophulariaceae and assuming a molecular clock, a date of 40 My for the split of the Solanaceae and Scrophulariaceae has been proposed (R. Olmstead, personal communication, in Xue *et al.* 1996). However it is not mentioned what fossil data were used to calibrate the molecular clock.

Rate calibration may be a much greater problem in molecular clock dating than unequal rates of sequence evolution (Bremer 2000). The fossil data are few for angiosperms and there are large variances associated with the dates. I have used the data given by Bremer (2000) on *rbc*L. This gives the highest and lowest values for the ratio of the branch lengths / estimated age of the fossil used to calibrate given splits between taxa. Based on these rates and the average synonymous site divergence between *Antirrhinum* / *Verbascum* and *Nicotiana* for the *rbc*L gene (Table 6.3) the minimum and maximum estimates for the age of the Solanaceae / Scrophulariaceae split are 42 and 122 My, respectively. The latter value is most likely an overestimate, as the date for the major diversification of angiosperms is between 130 and 90 MYA (Crane *et al.* 1995). In what follows, I use the minimum estimate of about 40 My for the Solanaceae / Scrophulariaceae split.

Table 6.3. Silent site divergence within the Scrophulariaceae and between the Scrophulariaceae and Solanaceae

Gene	Length of region analysed (bp)	Species compared		Silent site divergence
*ndh*F	2093	*A. majus*	*V. thapsus*	0.1751
		A. majus	*Nicotiana tabacum*	0.3504
		V. thapsus	*N. tabacum*	0.2814
*trn*L	518	*A. majus*	*V. nigrum*	0.0531
		A. majus	*N. tabacum*	0.0848
		V. nigrum	*N. tabacum*	0.0774
*rbc*L	1329	*A. majus*	*V. thapsus*	0.0865
		A. majus	*N. tabacum*	0.1914
		V. thapsus	*N. tabacum*	0.2018

Accession numbers for *ndh*F are: *A. majus* L36392, *V. thapsus* L36417, *N. tabacum* L14953; for *trn*L are: *A. majus* AF118790, *V. thapsus* AF118804, *N. tabacum* M16898; and for *rbc*L are: *A. majus* L11688, *V. thapsus* L36452, *N. tabacum* Z00044 (94073993).

Assuming a molecular clock (Zuckerkandl and Pauling 1965), I can then estimate the age of the *Antirrhinum* and *Verbascum* split. The silent substitution rate for three chloroplast genes is presented in Table 6.3. From these, the average estimated, divergence time between *Antirrhinum* and *Verbascum* is 22 My (22.2 My for *ndh*F; 26.2 My for

*trn*L; 17.6 My for *rbc*L). Very similar values are obtained when species of Solanaceae other than *Nicotiana* are used (data not shown). For the 18S rDNA nuclear sequences, however, the estimates depend on the species of Solanaceae used (Table 6.4), and range from 12 to 25 My (average 19 My). The large variance observed may be due to the limited number of differences in these comparisons, since this gene evolves slower than the three chloroplast genes analysed (Soltis *et al.* 1997). Overall, the data suggest a minimum age for the split between *Antirrhinum* and *Verbascum* of about 20 My, and rule out a recent date for the evolution of *Antirrhinum* and *Verbascum*. The *Adh* data presented here are also compatible with a non- recent origin of the Scrophulariaceae.

(ii) A low rate of nucleotide substitution

Another explanation that could account for the low divergence observed for *fil1*, *fil2*, *far*, and *globosa* genes between *Antirrhinum* and *Verbascum* is a low rate of nucleotide substitution. The estimated nucleotide substitution rates for the *cyc*-like genes (Chapter 4), *fil1A* genes (Chapter 5), and the four genes described here are lower than most other estimates for plant nuclear genes in monocotyledons (Wolfe *et al.* 1989; Gaut *et al.* 1996) and dicotyledons (Small *et al.* 1999; Small and Wendel 2000). The low levels of divergence are not an artifact of the uncertain age of the split between the species analysed. This rate is still lower than in other plant species, even assuming that divergence time is greatly overestimated.

Table 6. 4. Silent site divergence within and between Scrophulariaceae and Solanaceae for the 18S rDNA nuclear gene (1737 bp analysed)

Species compared	Silent site divergence	
	A. majus	*V. thapsus*
V. thapsus	0.0059	-
Other species names (Solanaceae)		
Alstonia scholaris	0.0202	0.0190
Nolana humifusa	0.0088	0.0100
Petunia axillaris	0.0148	0.0148
Nicotiana tabacum	0.0088	0.0100
Schizanthus sp.	0.0112	0.0124
Duckeodendron cestroides	0.0136	0.0148

Accession numbers are: *A. majus* AJ236047; *V. thapsus* AF207051/ AF16101; *A. scholaris* AF107570; *N. humifusa* AJ236017; *P. axillaris* AJ236020; *N. tabacum* AJ236016; *Schizanthus* sp. AF207016; *D. cestroides* AF206904

In conclusion since the observation of low divergence between *Antirrhinum* and *Verbascum* comes mainly from genes that are involved in flower development, it is possible that this is a feature of the structure of these genes, or of genes that belong to large gene families. This, however, seems not to be a general pattern since allozyme divergence is observed within and between species of Scrophulariaceae (Elisens and Crawford 1988; Hamrick and Godt 1990; Ritland 1989; Schoen and Brown 1991; Elisens

1992; Elisens and Nelson 1993; Mateu-Andres 1999). Our data on *Adh* also differ from the results for the other loci. More comparative studies at the DNA level using non-developmental genes are needed. It may be helpful to study genes for which allozyme divergence among species of Scrophulariaceae has been shown.

Chapter 7

Characterisation of 35 putative S-alleles from *Antirrhinum* and *Misopates* species

7.1 Introduction

Self-incompatibility in flowering plants is determined by a genetic system that prevents self-fertilization by enabling the pistil to reject pollen from genetically related individuals, and is controlled by a single multi-allelic locus called the S-locus (de Nettancourt 1977; see Chapter 1). In Scrophulariaceae, Solanaceae and Rosaceae, which have gametophytic self-incompatibility systems (GSI), the S-locus products are basic glycoproteins with ribonuclease activity (e.g. Matton *et al.* 1994, Golz *et al.* 1995, Kao and McCubbin 1996).

In *Antirrhinum* (Scrophulariaceae) the sequences of three S-alleles were obtained by Xue *et al.* (1996) from crosses of *Antirrhinum majus* strains used in the horticulture industry (that are self compatible and have non-functional S-alleles; Baur 1919; Sherman 1939) with three genotypically distinct lines of *A. hispanicum* (which are self-incompatible). F1 plants (heterozygous for a non-functional and a functional S- allele)

were then intercrossed to obtain some self-incompatible (SI) F2 plants. Because another gene unlinked to the S-locus can cause self-compatibility in *A. majus* (Brieger 1935; Tseng 1938), in order to obtain a population where all individuals are SI, crosses with two F2 individuals were performed. Four groups of plants were identified by reciprocal crosses between individuals. Crosses between individuals of different groups, were always compatible. Plants from the F3 were used to isolate three cDNAs encoding polypeptides homologous to S-RNases (Xue *et al.* 1996). Analysis of these three *Antirrhinum* S-alleles indicated four highly conserved short stretches (named C1, C2, C3 and C5) and two hypervariable regions (named HVa and HVb; Xue *et al.* 1996). These regions are also found in S-RNases of Solanaceae (Ioerger *et al.* 1990; Tsai *et al.* 1992). In Rosaceae, these regions but the HVb are also conserved (Norrioka *et al.* 1996; Sassa *et al.* 1996; Ushijima *et al.* 1998). The conserved regions contain mostly hydrophobic amino acids and are thought to be involved in the core structure of the S protein. The C2 and C3 regions are also found in RNases of *Rhizopus niveus* (a fungus), and contain two of the three histidine residues required for catalytic activity (Ohgi *et al.* 1992).

A number of genes with similar structure to S-RNases but not linked to the S-locus (the "S-like RNases") have also been described in various plant families including Solanaceae (Jost *et al.* 1991; Löffler *et al.* 1992; Löffler *et al.* 1993; Lee *et al.* 1992; Kuroda *et al.* 1994) and Rosaceae (Norioka *et al.* 1996). Expression of S-like RNase genes has been associated with phosphate starvation (Löffler *et al.* 1992; Bariola *et al.* 1994; Dodds *et al.* 1996), leaf senescence (Lers *et al.* 1998) and wounding (Ye and Droste 1996). They are encoded at a number of separate genetic loci and their evolution is characterised by gene duplications and slow accumulation of sequence differences (Ma

and Oliveira 2000). Therefore Southern hybridisation and PCR amplification experiments always detect one band in every individual, corresponding to the specific S-like RNase being studied (Lee *et al.* 1992, Dodds *et al.* 1996, Norioka *et al.* 1996). This is in contrast with the S-alleles for which, even under low stringency conditions, a band is not obtained for every individual by Southern hybridisation and/ or PCR amplification, due to their high sequence divergence.

The HVa and HVb regions are the most hydrophilic regions of the S-proteins (Ioerger *et al.* 1991), and therefore they may lie on the external surface of the protein and could play an important role in the discrimination of self from non-self pollen (Ioerger *et al.* 1991; Saba-El-Leil *et al.* 1994; Matton *et al.* 1997). With the exception of four sequences of *Prunus avium* (Rosaceae), which have at least two introns (Tao *et al.* 1999), the S genes usually have a single intron in the HVa region of species of Solanaceae and Rosaceae. This can vary from 87 bp to 120 bp in Solanaceae (Saba-El-Leil *et al.* 1994; Matton *et al.* 1995) and 138 bp to 1100 bp in Rosaceae (Broothaerts *et al.* 1995). The intron is mostly, but not always, in the same position relative to the flanking coding sequence (Ebert *et al.* 1989; Jahnen *et al.* 1989; Kaufmann *et al.* 1991; Tsai *et al.* 1992; Coleman and Kao 1992; Saba-El-Leil *et al.* 1994). In some alleles the intron can be translated (Saba-El-Leil *et al.* 1994). It has therefore been proposed that by shifting the intron borders new S proteins could arise and potentially produce proteins of different specificity (Saba-El-Leil *et al.* 1994).

Two different types of molecular experiments have been used to test the importance of regions other than the HV in determining specificity. Swapping entire regions between very different Solanaceae S-alleles led to the conclusion that the

recognition function of the S-RNase is not localised to a specific domain in any of the RNases studied, and that HV regions are necessary but not sufficient for encoding specificity (Kao and McCubbin 1997; Zurek *et al.* 1997). However, Matton *et al.* (1997) showed that replacing the hypervariable region of one S-allele from *S. chacoense* with that from another with a very similar sequence caused the appropriate change in the pistil incompatibility. They therefore concluded that the HV regions control allelic specificity. Later, Matton *et al.* (1999) showed that a single amino acid change in the HVa region of an S-allele from *S. chacoense* produces a new S-allele with dual-specificity incompatibility, that simultaneously rejects two phenotypically and genotypically distinct types of pollen tubes.

The sequences currently available do not include any independently isolated alleles having the same incompatibility type. In the only available comparison, in *P. rhoeas* (Papaveraceae), only silent site differences were found (Walker *et al.* 1996). The availability of such data could help provide evidence about which parts of the gene are not involved in determining specificity differences.

The frequency of amino acid and silent changes in different regions of the gene may also help identify regions that are important for specificity determination. Recombination allows different segments of a gene to have different evolutionary histories (Hudson 1983, 1990). Therefore silent sites linked to the target of balancing selection are expected to have higher polymorphism levels than elsewhere in the gene. However in the absence of recombination there should be no variation other than stochastic variability along the gene in silent variability levels (Hudson 1990), and this approach, therefore, would not be informative. Since S-alleles are expected to persist in a

population for long evolutionary periods of time (reviewed by Clark 1993), even in the presence of recombination it may be difficult to detect regions that are important for specificity determination because of saturation at silent sites (Hinata *et al.* 1995, Uyenoyama 1997).

K_a/K_s ratios ~1 are usually considered evidence of neutral evolution (Nei 1987), whereas genes under balancing selection may have values of this ratio >1 (Hughes *et al.* 1990). In *Solanum* HV regions, it has been shown that the average ratio of non-synonymous to synonymous changes is greater than one between pairs of alleles that are most closely related. This pattern suggests that selection is favouring new allelic types differing in amino acid sequence (Richman *et al.* 1996b, Charlesworth and Guttman 1997, Richman and Kohn 1999). It is not known, however, which amino acid variants in these regions are maintained by balancing selection.

I have determined and analysed the genomic sequence of 35 putative S-alleles from *Antirrhinum* and *Misopates* species. The region under study includes conserved regions C2 and C3, both HV regions (HVa and HVb) and the intron region.

7.2. Material and Methods

Plant material and DNA extraction

Genomic DNA of leaves of individual plants collected in the field was extracted using the method of Ingram *et al.* (1997), from three self-incompatible *Antirrhinum* species *(A. graniticum*, *A. molle* and *A. meonanthum*), two largely self-compatible *A. majus* subsp. *cirrhigerum* populations (cirrhigerumAve and cirrhigerumGala), one largely self-incompatible *A. majus* subsp. *cirrhigerum* (cirrhigerumMuel), one self-compatible

Misopates orontium population and one self-compatible *Misopates calycinum* population (Table 3.1 and Table 3.2 in Chapter 3).

A. graniticum crosses

Two plants (graniticumB1-1 and graniticumB1-2) were grown from two seeds of a fruit of a wild individual of *A. graniticum* (graniticumB1). Crosses between these two plants, in both directions, produce fruits with many seeds. Three other plants (graniticumB4-1, graniticumB4-2, and graniticumB4-3) were grown from three seeds from another wild plant (graniticumB4) of the same population. Crosses between these three plants also produce fruits with many seeds. Four crosses were then performed: graniticumB1-1 with graniticumB4-1, graniticumB1-1 with graniticumB4-2, graniticumB1-1 with graniticumB4-3, and graniticumB1-2 with graniticumB4-3. All crosses produced fruits with many seeds. A total of 21 plants from seeds of these crosses were used to make cross-pollinations in both directions in a pair-wise fashion (Table 7.1). Each pollination was repeated usually three times. If fruits containing seeds were obtained, the plants must differ in at least one S allele. However, if no seeds were obtained, the plants must have the same S genotype, and are included in the same group. Unfortunately, I could not perform segregation analyses since the information on the parents of the 21 plants was lost as a result of mislabelling.

PCR amplification

Based on the GenBank sequences of S2, S4 and S5 *Antirrhinum* S-alleles (accession numbers X96465, X96466, and X96464 respectively), primer ants2s3F (position 108 and 112 in relation to the start codon of the S2 and S5 cDNA sequence) and ants4F

Table 7.1. *Antirrhinum graniticum* pollinations in the glasshouse. Number of pollinations which produced seeds / total number of pollinations. Dashes indicate pollinations not performed. In bold are shown the result of pollinations within specificity groups.

	Compatibility groups																				
	A					B					C		D			E				F	G
Plant code	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	**0 / 5**	**0 / 3**	**0 / 2**	**0 / 6**	**0 / 2**	3 / 3	2 / 2	-	2 / 2	2 / 2	6 / 6	4 / 4	3 / 3	3 / 3	4 / 4	3 / 3	2 / 3	2 / 2	3 / 3	3 / 3	3 / 3
2	**0 / 3**	**0 / 5**	**0 / 3**	**0 / 6**	**0 / 6**	3 / 3	3 / 3	-	3 / 3	3 / 3	3 / 3	2 / 3	3 / 3	3 / 3	-	3 / 3	-	3 / 3	2 / 3	3 / 3	-
3	**-**	**0 /3**	**0 / 5**	**0 / 3**	**0 / 3**	-	-	-	-	-	-	-	-	-	-	-	-	3 / 3	-	-	-
4	**0 / 5**	**0 / 3**	**0 / 3**	**0 / 5**	**0 / 3**	3 / 3	-	-	2 / 2	2 / 2	-	3 / 3	3 / 3	3 / 3	3 / 3	-		3 / 3	3 / 3	3 / 3	-
5	**0 / 3**	**-**	**0 / 3**	**0 / 5**	**0 / 5**	2 / 3	2 / 3	2 / 3	-	-	3 / 3	3 / 3		2 / 2	-	-	-	-	-	3 / 3	-
6	3 / 3	2 / 2	2 / 2	3 / 3	-	**0 / 5**	**2 / 7**	**-**	**0 / 2**	**0 / 3**	-	2 / 2	3 / 3	3 / 3	3 / 3	3 / 3	2 / 2	1 / 1	3 / 3	2 / 2	3 / 3
7	3 / 3	2 / 2	2 / 2	-	3 / 3	**0 / 3**	**0 / 5**	**0 / 3**	**0 / 2**	**0 / 2**	3 / 3	-	-	3 / 3	-	3 / 3	3 / 3	3 / 3	3 / 3	3 / 3	3 / 3
8	-	2 / 2	2 / 2	-	-	**-**	**0 / 3**	**0 / 5**	**0 / 2**	**0 / 2**	3 / 3	3 / 3	-	-	-	3 / 3	3/3	3 / 3	3 / 3	3 / 3	-
9	3 / 3	3 / 3	-	3 / 3	-	**0 / 3**	**0 / 3**	**0 / 3**	**0 / 5**	**0 / 3**	3 / 3	-	-	-	-	2 / 2	-	3 / 3	3 / 3	-	3 / 3
10	3 / 3	2 / 3	-	-	-	**0 / 3**	**-**	**0 / 3**	**0 / 4**	**0 / 5**	3 / 3	-	-	-	-	4/ 6	-	3 / 3	3 / 3	-	-
11	3 / 3	2 / 2	2 / 2	3 / 3	3 / 3	-	3 / 3	2 / 3	-	-	**0 / 5**	**0 / 3**	-	2 / 2	-	-	-	-	-	-	-
12	3 / 3	-	-	-	3 / 3	-	-	-	-	-	**0 / 3**	**0 / 5**	-	-	-	3 / 3	-	-	-	-	-
13	3 / 3	2 / 2	2 / 2	3/3	-	3 / 3	-	-	2 / 2	2 / 2	-	2 / 2	**0 / 5**	**0 / 3**	**0 / 3**	3 / 3	2 / 3	3 / 3	3 / 3	6 / 6	3 / 3
14	3 / 3	2 / 2	2 / 2	6 / 6	3 / 3	3 / 3	3 / 3	3 / 3	2 / 2	2 / 2	3 / 3	-	**0 / 8**	**0 / 5**	**0 / 5**	3 / 3	3 / 3	-	3 / 3	6 / 6	3 / 3
15	2 / 3	-	-	3 / 3	-	3 / 3	-	-	2 / 2	3 / 3	-	-	**0 / 3**	**0 / 3**	**0 / 5**	-	3 / 3	-	2 / 3	-	-
16	-	-	-	-	2 / 2	-	-	-	3 / 3	3 / 3	3 / 3	-	-	-	-	**0 / 5**	**0 / 3**	**0 / 3**	**0 / 3**	-	3 / 3
17	-	-	-	-	-	-	-	-	-	-	3 / 3	-	-	-	-	**0 / 3**	**0 / 5**	**0 / 3**	**0 / 2**	-	-
18	-	3 / 3	3 / 3	-	-	3 / 3	2 / 2	2 / 2	3 / 3	-	-	3 / 3	-	3 / 3	-	**0 / 3**	**0 / 3**	**0 / 5**	**0 / 3**	2/ 2	3 / 3
19	3 / 3	-	-	3 / 3	3 / 3	3 / 3	2 / 3	3 / 3	3 / 3	2 / 3	3 / 3	3 / 3	2 / 2	3 / 3	-	**0 / 3**	**-**	**1 / 5**	**0 / 5**	3 / 3	-
20	2 / 2	-	-	3 / 3	3 / 3	-	-	-	3 / 3	-	3 / 3	3 / 3	3 / 3	2 / 3	-	0 / 3	-	3 / 3	2 / 3	**0 / 5**	4 / 4
21	4/ 5	-	-	3 / 3	-	3 / 3	3 / 3	-	-	-	3 / 3	3 / 3	-	-	-	2 / 2		-	-	-	**0 / 5**

(position 109 in relation to the start codon of the S4 cDNA sequence) were designed for the conserved region C1 (Table 7.2). Primer ants2s3R (position 560 in relation to the start codon of the S2 and S5 cDNA sequence) and ants4R (position 560 in relation to the start codon of the S4 cDNA sequence) were designed for the conserved motif LKC 5' to the conserved region C5. Because of the size of the S4 region analysed, DNA double sequence cannot be obtained for its genomic region without additional internal primers. Therefore two additional primers were designed (ants4-1 designed based on the sequences here obtained, and ants4-2 designed for position 272 in relation to the start codon of the S4 cDNA sequence). Standard amplification conditions were 35 cycles of denaturation at 94° C for 30 seconds, primer annealing at 49° C for 30 seconds, and primer extension at 72° C for 2 minutes.

Table 7.2. Primers used in this study

Forward primers	
ants2s3F	5' AAATTGGTTCT(TC)CAATGG 3'
ants4F	5' AAGCTAGTTCTCCAATGG 3'
ants4-1	5' GATTGCTCAGAACTTTTTAC 3'
Reverse primers	
ants2s3R	5' AACCTTCA(CG)(CG)GCATTTCA 3'
ants4R	5' ATAACTTTGAGCACATTTCA 3'
ants4-2	5' TCTCTGCTCTATGTCTGTAT 3'

I routinely checked the PCR amplification products for homogeneity by digestion with several four-cutter restriction enzymes. If the number and/or the size of the bands obtained after digestion is not compatible with that of the reference sequence (from which the primers were designed), the amplification product was classified as heterogeneous. In such cases, I cloned the product and screened several colonies until several of each of the types previously revealed were found, and determined their DNA sequences. Cloning was performed using the TA cloning kit (Invitrogen). If more than one band was systematically obtained, I always cloned and sequenced them. Because differences can arise from nucleotide mis-incorporation during amplification, we determined the DNA sequence of at least three different colonies and obtained a consensus sequence. DNA sequencing was performed with an Applied Biosystems model 377 DNA sequencing system with the ABI PRISM BigDye cycle-sequencing Kit (Perkin Elmer), using specific primers or the primers for the M13 forward and M13 reverse priming sites of the pCR 2.1 vector.

Analyses of the sequences

The nucleotide sequences of the putative coding regions to be compared were aligned using ClustalX v. 1.64b (Thompson *et al.* 1997), and minor manual adjustments were performed using SeqPup v. 0.6f. The numbers of synonymous and nonsynonynous differences between pairs of sequences were calculated using the DnaSP software (Rozas and Rozas 1997). Neighbor-joining trees were generated with MEGA version 1.01 (Kumar *et al.* 1994).

7.3. Results

7.3.1. Analysis of A. graniticum crosses

The presence/absence of seeds from the crosses between 21 *A. graniticum* plants (see Material and Methods) allowed me to divide these plants into seven self-incompatible groups (A - G; Table 7.1 and Table 7.3). All groups except F and G have more than one individual. Within groups B (plants 6 and 7) and E (plants 18 and 19) in two and one pollinations, respectively, seeds were obtained. This is unexpected since individuals within groups are inferred to have the same S-locus genotype and therefore are fully self-incompatible.

Table 7.3. Summary of the pollinations in Table 7.1. Number of pollinations which produced seeds / total number of pollinations. Dashes indicate pollinations not performed.

Compatibility groups	Number of plants	A	B	C	D	E	F	G
A	5	0 / 65	34 / 37	24 / 27	27 / 27	27 / 29	12 / 12	3 / 3
B	5	38 / 39	2 / 51	17 / 17	12 / 12	51 / 53	8 / 8	9 /9
C	2	19 / 19	5 / 6	0 / 6	2 / 2	3 / 3	-	-
D	3	31 / 32	28 / 28	5 / 5	0 / 25	25 / 27	12 / 12	6 / 6
E	4	17 / 17	29 / 31	15 / 15	8 / 8	1 / 34	5 / 5	6 / 6
F	1	8 / 8	3 / 3	6 / 6	5 / 6	5 / 6		4 / 4
G	1	7 / 8	6 / 6	6 / 6	-	-	-	

These are likely to be the result of rare contamination with pollen from other plants. In 3.6% (20 cases out of 548; Table 7.3) of the pollinations between different groups no

seeds were obtained. These are likely to be the result of damage during handling of the flowers used in pollinations.

Genomic DNA of these 21 individuals was used in PCR amplifications using two different pairs of primers (ants2s3F, ants2s3R that were designed based on the S2 and S3 *Antirrhinum* sequences, and ants4F, ants4R, designed based on the S4 *Antirrhinum* sequence; see Material and Methods).

The primer pair ants4F, ants4R did not amplify any of the S-alleles present in these 21 plants. Using primers ants2s3F and ants2s3R, a 658 bp long product was obtained for all eight individuals of three groups, B, C and G. From digestions of the amplification products with several restriction enzymes (*Aci*I; *Alu*I, *Rsa*I, and *Dde*I) I conclude that they are homogeneous. The restriction enzyme pattern is identical among the eight individuals, and the sequence was obtained for one individual. The highest amino acid similarity by BlastX search (58 % amino acid identity) is with the published *Antirrhinum* S5-allele sequence. There is an insertion of 186 bp in our sequence relative to the *Antirrhinum* S5 S-allele. Since the *Antirrhinum* S-allele sequences in GenBank are cDNAs, this is likely to be an intron sequence. The putative intron/exon boundaries (5' splice site CAA : GTATAG; 3' splice site TGATGCATGGTTCGCCAG : GA) are in agreement with the deduced consensus splice sites that we have found for 31 introns of *Antirrhinum* nuclear genes (Table 7.4) and *Arabidopsis thaliana* (Brown *et al.* 1996). Therefore this sequence seems to be a new functional *Antirrhinum* S-allele, and I named it S6 (since Xue *et al.* (1996) have used the names S1-S5 to describe *Antirrhinum* S-alleles).

Table 7.4. 5' and 3' splice sites for 31 *Antirrhinum* introns and consensus sequence.

Gene	5' splice site									3' splice site																				
chs (X03710)																														
Intron 1	T	G	T	G	T	A	A	G	A	T	T	T	A	A	T	T	T	G	A	A	T	T	A	T	C	A	G	G	C
Intron 2	C	A	G	G	T	A	C	G	T	T	T	T	A	T	T	A	T	C	C	A	A	C	A	C	T	A	G	G	G
TAP1 (X57295)																														
Intron 1	A	T	G	G	T	A	T	A	C	T	A	A	C	A	T	T	T	G	T	T	G	T	T	G	C	A	G	A	T
DEFA (X62810)																														
Intron 1	T	G	C	G	T	A	C	G	T	T	T	G	T	T	T	G	T	T	T	T	T	A	A	A	C	A	G	G	A
Intron 2	G	A	G	G	T	T	T	C	C	C	G	G	T	G	A	A	A	A	T	A	T	G	T	G	C	A	G	A	A
Intron 3	C	A	G	G	T	A	C	A	T	T	A	A	C	T	T	T	A	T	A	A	T	G	T	G	C	A	G	G	C
Intron 4	A	A	G	G	T	T	T	T	A	A	C	G	C	T	T	C	G	G	T	C	G	T	A	A	C	A	G	T	A
Intron 5	A	A	G	G	T	A	T	G	T	C	T	T	T	G	T	T	T	A	T	T	T	A	T	G	C	A	G	G	T
Intron 6	T	T	T	G	T	A	A	G	C	A	T	G	A	T	G	T	T	T	A	A	T	T	C	G	T	A	G	G	A
fil2 (X76995)																														
Intron 1	C	A	G	G	T	C	A	G	T	A	A	G	T	A	A	G	A	T	A	T	T	T	T	T	C	A	G	G	C
far (AJ239057)																														
Intron 1	C	A	G	G	T	A	A	A	A	A	T	T	A	A	T	G	C	A	A	A	T	G	T	G	C	A	G	T	G
Intron 2	C	A	G	G	T	A	C	C	T	C	C	T	C	A	G	C	A	A	T	T	A	T	T	T	C	A	G	T	A
Intron 3	C	A	G	G	T	A	A	T	T	G	T	T	C	T	G	T	T	C	T	T	G	A	T	T	T	A	G	G	A
Intron 4	A	A	G	G	T	A	A	T	T	C	T	T	C	A	T	T	A	T	A	T	G	C	C	A	C	A	G	A	A
Intron 5	A	G	G	G	T	A	A	G	T	G	G	G	G	T	A	A	A	A	T	A	T	T	G	G	C	A	G	C	A
Intron 6	A	A	G	G	T	C	A	T	T	T	T	T	T	T	G	T	A	T	A	T	T	G	T	G	T	A	G	A	T
Intron 7	A	G	T	G	T	A	A	G	T	T	T	T	T	T	T	T	C	T	C	A	A	T	T	A	C	A	G	C	T
squamosa (X63701)																														
Intron1	T	T	G	G	T	T	A	G	T	A	T	A	A	T	G	G	T	T	C	A	T	A	A	G	C	A	G	C	A
Intron 2	C	C	T	G	T	A	A	G	T	T	T	T	T	A	T	G	A	T	T	T	T	G	T	G	T	A	G	G	C
Intron 3	T	A	G	G	T	A	A	A	T	C	T	T	A	A	T	T	T	C	T	T	C	T	G	T	C	A	G	G	C
Intron 4	A	A	A	G	T	A	T	T	A	T	C	G	T	A	T	G	A	T	C	C	A	A	T	G	C	A	G	A	A
Intron 5	A	A	G	G	T	T	A	G	T	A	A	T	C	T	G	T	T	T	T	T	T	T	G	A	C	A	G	G	A
Intron 6	A	A	G	G	T	A	A	C	T	T	A	T	T	G	T	T	C	T	G	T	G	A	T	A	T	A	G	A	T
Intron 7	C	G	G	G	T	A	A	T	T	A	G	A	G	G	T	C	T	A	T	T	G	A	T	G	C	A	G	A	A
fil1 (X57296)																														
Intron 1	G	T	G	G	T	G	A	G	C	A	A	T	A	T	T	T	T	T	T	T	C	C	T	T	G	C	A	G	C	A
globosa (X68831)																														
Intron 1	C	A	C	G	T	A	T	G	T	T	T	T	T	G	T	T	T	G	A	T	A	A	A	T	T	A	G	G	T
Intron 2	G	A	G	G	T	A	A	A	T	A	A	T	A	A	C	G	T	T	A	T	T	T	A	A	C	A	G	C	A
Intron 3	C	A	G	G	T	A	A	A	A	C	T	A	C	T	T	G	T	G	A	A	T	T	C	T	A	A	G	G	C
Intron 4	C	A	G	G	T	T	C	A	T	G	T	T	T	T	T	G	T	T	T	T	G	T	G	G	C	A	G	A	T
Intron 5	C	A	T	G	T	A	T	A	A	A	T	T	C	T	T	T	T	C	T	T	T	T	T	T	T	A	G	A	A
Intron 6	C	T	G	G	T	A	C	G	T	T	T	G	C	G	A	T	A	T	T	G	C	T	A	A	C	A	G	C	G
Consensus sequence*	C_{42}	A_{65}	G_{74}	G_{100}	T_{100}	A_{74}	A_{58}	G_{45}	T_{68}	T_{39}	T_{58}	T_{58}	T_{35}	T_{48}	T_{65}	T_{52}	T_{55}	T_{52}	T_{52}	T_{55}	T_{52}	T_{52}	T_{52}	G_{45}	C_{71}	A_{100}	G_{100}	G_{42}	A_{48}

* Percentage occurrence of the consensus nucleotides are given in subscript. For the 5' splice site the sequence 5' to the 100% conserved GT is exon sequence. For the 3' splice site the sequence 3' to the 100% conserved AG is exon sequence.

7.3.2. Distribution of putative S-alleles in plants sampled from natural populations

Primers ants2s3F and ants2s3R were used in PCR amplifications of genomic DNA of four *Antirrhinum* and two *Misopates* species. From a total of 170 sampled in the wild, from different populations and species, products were amplified from 30 individuals (Table 7.5). The size of the PCR amplification product varied from 604 bp (graniticumB15 sequence) to 689 bp (cirrhigerumMuel7, molle2, molle3, molle10, molle13, molle16, molle17, graniticumGala9, orontiumB26 and orontiumB27 sequences). When these were digested with the restriction enzymes as above, only the amplification product of the molle4 individual was heterogeneous. This individual's PCR amplification product was cloned. Twenty clones were digested using the same restriction enzymes as above, revealing only two types of clones (A has two *Aci*I and one *Rsa*I restriction sites; B has one *Aci*I and no *Rsa*I restriction sites). Several clones of each type were sequenced.

Three (graniticumB4, graniticumB6 and graniticumB29) of the total 31 sequences obtained were identical to the S6 allele present in the eight *A. graniticum* plants used in the glasshouse pollinations (see above). Another individual (molle4, cloneA) differs from S6 sequences by a one bp indel in the putative intron region. BlastX search revealed that the remaining 28 sequences share the highest amino acid similarity (ranging from 28% to 98% amino acid identity) with the published *Antirrhinum* S2 or S5-allele sequences. A putative intron was deduced in all the sequences by comparison with the S-allele cDNA sequences in GenBank. Intron/exon boundaries are conserved among these sequences (Table 7.6) and are again in agreement with the deduced consensus splicing sites found for *Antirrhinum* nuclear genes (Table 7.4). The sizes of the putative intron varies from

Table 7.5. Summary of the species, population codes, mating system and individuals which supported the PCR amplification with two primer sets.

Species and population code	Mating system	N	Individuals that yield PCR product with primers	
			ants2s3F+ants2s3R	ants4F+ants4R
A. majus ssp. cirrhigerum				
cirrhigrerumAve	largely self-compatible	21		
cirrhigerumGala	largely self-compatible	19	cirrhigerumGala5 (B) cirrhigerumGala7 (B) cirrhigerumGala9 (C) cirrhigerumGala18 (B)	cirrhigerumGala9 (S4) cirrhigerumGala 11 (S4) cirrhigerumGala 14 (S4)
cirrhigerumMuel	largely self-incompatible	13	cirrhigerumMuel7 (C)	
A. majus ssp. linkianum				
linkianum	partially self-incompatible	15	linkianum1 (S5)	
A. graniticum				
graniticumB	self-incompatible	28	graniticumB4 (S6) graniticumB5 (A) graniticumB6 (S6) graniticumB15 (unique) graniticumB24 (C) graniticumB27 (unique) graniticumB29 (S6)	graniticumB14 (S4)
graniticumS	not determined	17	graniticumS5 (S6)	
A. molle				
molle	self-incompatible	18	molle2 (C) molle3 (C) molle4 (cloneA- S6; cloneB- A) molle5 (A) molle9 (A) molle10 (C) molle13 (C) molle16 (C) molle17 (C)	
A. meonanthum				
meonanthum	self-incompatible	7	meonanthum1 (A) meonanthum3 (A) meonanthum4 (unique) meonanthum7 (B) meonanthum8 (B)	
Misopates orontium				
orontiumB	self-compatible	22	orontiumB27 (C) orontiumB26 (C)	
Misopates calycinum				
calycinum	self-compatible	10		

N is the number of individuals tested; the sequence groups (in brackets) were defined according to levels of divergence between sequences (see text for details).

132 to 217 bp (see below and Fig. 7.1).

Primers ants4F and ants4R, designed based on the cDNA S4 *Antirrhinum* sequence (see Material and Methods), amplified a PCR product in only four individuals of two *Antirrhinum* species out of 170 individuals tested (Table 7.5). The four PCR amplification products are of similar size (1386 bp in graniticumB14 and 1389 bp in cirrhigerumGala9, cirrhigerumGala11, and cirrhigerumGala14). Digestion with the restriction enzymes mentioned above suggested that the four amplification products are homogeneous. Sequences were obtained for these amplification products using primers ants4F, ants4-1, ants4-2, and ants4R (see Material and Methods). BlastX search revealed that the four sequences share the highest amino acid similarity (> 97% amino acid identity) with the published S4 *Antirrhinum* sequence. Comparison with the S4 cDNA sequence in GenBank revealed the presence of a putative intron 953 bp long in the graniticumB14 sequence, and 956 bp long in the cirrhigerumGala9, cirrhigerumGala11 and cirrhigerumGala14 sequences. Intron/exon boundaries are conserved among these sequences (Table 7.6) and are again in agreement with the deduced consensus splice sites found for *Antirrhinum* nuclear genes (Table 7.4).

All the *Antirrhinum* and *Misopates* putative intron sequences when included in the ORF contain several stop codons. Blast searches of these sequences reveals no similarity with any sequence in GenBank.

Table 7.6. The splicing sites of 35 *Antirrhinum* putative S-alleles.

	5' Splice site	3' Splice site
molle4 graniticumB4 graniticumB6 graniticumB29	**C A A G** T A T A G	T G A T G C A T G G T T C G C C A G **G A**
graniticumB5 graniticumS5 meonanthum3 molle4 molle5 molle9	**G A A G** T A T A G	T G A T G C A T T C C T G G C C A G **G A**
cirrhigerumGala5 cirrhigerumGala7 cirrhigerumGala18 meonanthum7 meonanthum8	**G A A G** T A T A A	T T G T G C A T G C C T C G C C A G **G A**
graniticumB24 cirrhigerumGala9 cirrhigerumMuel7 molle2 molle3 molle10 molle13 molle16 molle17 orontiumB26 orontiumB27	**A A A G** T A T G G	T G A C G C A T C A C T G A C C A G **G A**
linkianum1	**A C G G** T A C A T	A A T T A A T T G T G A T G C C A G **G A**
graniticumB14	**A C T G** T A A G T	T A A C C C A T G C C T A C G C A G **G A**
cirrhigerumGala9 cirrhigerumGala11 cirrhigerumGala14	**A C T G** T A A G T	T A A C G C A T G C C T A C G C A G **G A**
meonanthumR4	**G A A G** T A T G G	A G A T G C A T T A A C G T G C A G **G A**
graniticumB27	**G A G G** T A T G C	A T G G A T A T T C C T G C C C A G **A A**
graniticumB15	**G T T G** T A T A G	A T T A A C T T G T A T A C G C A G **C A**

In bold are exon sequences

The relationships among the 33 *Antirrhinum* and two *Misopates* sequences obtained here, and the three published *Antirrhinum* S-alleles (Xue *et al.* 1996) based on the coding region are shown in Figure 7.1. For two individuals (molle4 and cirrhigerumGala9) both S-alleles were obtained. Seven sequence groups (S6, A, B, S5, C, and S4) plus three unique sequences (graniticumB27, meonanthum4, S2, and graniticumB15) are apparent. Sequences from different species are included in the same group (Fig. 7.1). The S6 group contains the S-allele that is associated with three self-incompatibility groups of *A. graniticum* in the glasshouse. The published S4 *Antirrhinum* sequence cluster with the graniticumB14, cirrhigerumGala9, cirrhigerumGala11, and cirrhigerumGala14 sequences (group S4 in Fig 7.1). In the coding region, the sequences of group S4 differ in three nucleotide sites. Two are fixed between the four sequences obtained here and the published S4 *Antirrhinum* sequence (position 145 and 342, in relation to the start codon of *Antirrhinum* S4 sequence); the other difference is between graniticumB14 and the S4 *Antirrhinum* sequence (position 323). The *Antirrhinum* S5 sequence clusters with the linkianum1 sequence, which differs from it by two positions in the coding region (position 273 and 545 in relation to the start codon of *Antirrhinum* S5 sequence).

7.3.3. The conserved structure

The sequences here obtained have the structure expected for the region analysed, assuming that they are S-RNases. Figure 7.2 shows the pattern of variability, indicating the conserved regions C2 and C3 present in Solanaceae and Rosaceae. Only one sequence of each group is represented since within each group amino acid sequences are similar.

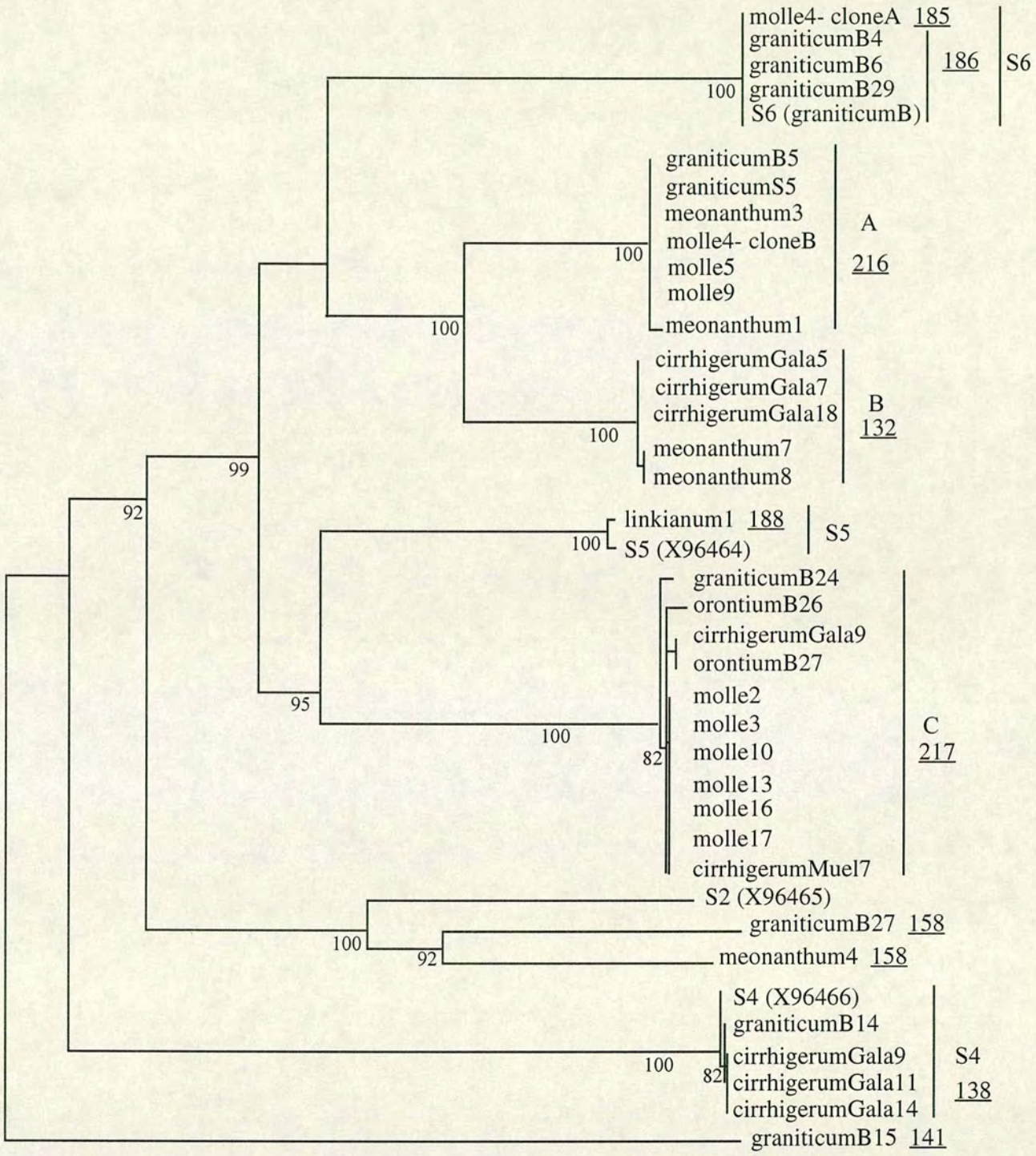


Fig. 7.1. Neighbor-joining tree using the Kimura two-parameter distance, showing the relationships among *Antirrhinum* and *Misopates* putative S-alleles and S2, S4 and S5 *Antirrhinum* S-alleles (Xue *et al*. 1996) coding sequences. Bootstrap replicates supporting the branches are shown for values greater than 80%. Intron sizes in bp are shown underlined.

The putative intron is at the same position in all sequences analysed (between amino acids 49 and 50 in Fig. 7.2). The C4 regions of Solanaceae (corresponding to amino acid 108 to 115 in Fig. 7.2) and Rosaceae (corresponding to amino acid 113 to 125 in Fig. 7.2) are not conserved regions in the *Antirrhinum* sequences (Xue *et al.* 1996). The hypervariable regions HVa and HVb are the most variable regions in the *Antirrhinum* and *Misopates* sequences (Fig. 7.2). Even when conserved regions C1, C2 and C3 are excluded from the analyses, the HV regions are more variable than the remaining sites (Mann-Whitney U test statistic; $P < 0.05$). The intron, which is located in this region has different sizes (Fig. 7.1), and therefore the alignment is ambiguous in this region. A total of 30 amino acid residues are conserved among all sequences (sixteen outside the conserved regions; Fig. 7.2). These are very likely to be functionally important amino acids. Four of these sites (12, 62, 81, and 111) are also conserved in the sequences from Solanaceae and Rosaceae, two (amino acids 102 and 115) in Solanaceae and two in Rosaceae (amino acids 63 and 80) according to Richman *et al.* (1997) and Ushijima *et al.* (1998). However, S-like RNases also share the conserved domains (Taylor *et al.* 1993; Sassa *et al.* 1996, Ushjima *et al.* 1998), and identification of S-alleles based on the structure of the gene alone is not possible.

7.4. Discussion

7.4.1. Evidence that the sequences studied are S-alleles

The sequences here obtained have the typical structure of S-RNases. However this structure is also shared by the S-like RNases. If the sequences here obtained were not linked to the S-locus, an amplification product should be obtained in all the

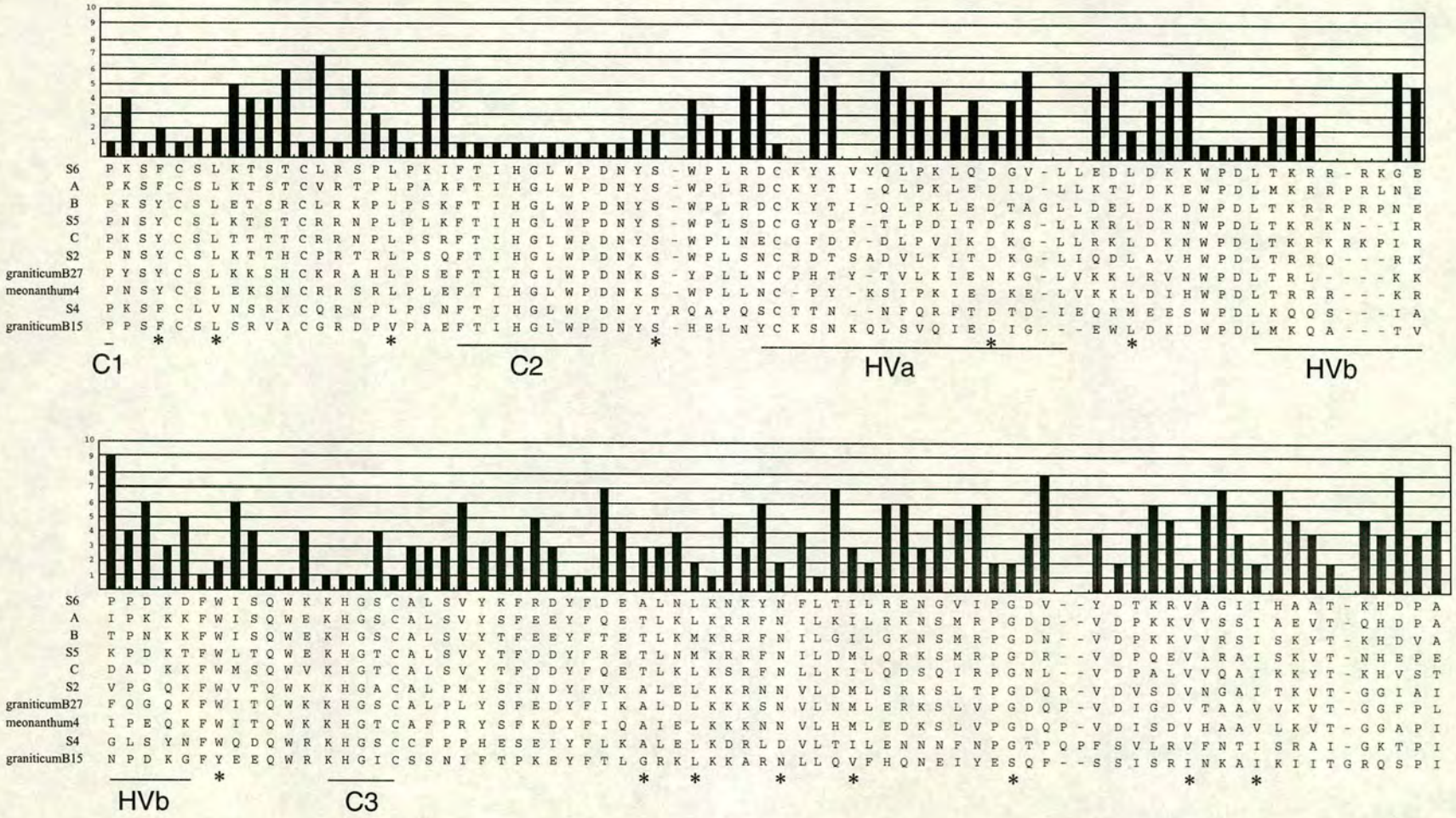


Fig 7.2. Amino acid sequences alignment of the *Antirrhinum* and *Misopates* sequences groups defined in Fig. 7.1. Bars indicate the numbers of different amino acids observed at each position. C1, C2 and C3 are conserved regions and HVa and HVb are hypervariable regions (Xue *et al.* 1996). Stars indicate conservative amino acid replacements (defined as in Dayhoff et *al.* 1979).

individuals of the same population and / or species. This is not the case (Table 7.5), I was able to amplify similar sequences from different species, but not from all individuals of a population. In addition some of the sequences obtained here are very similar to the S4 and S5 *Antirrhinum* S-alleles (Xue *et al.* 1996), and therefore they are also likely to be S-alleles. Furthermore, group S6 of sequences are very similar to the S6 sequence that has been shown to be associated with specificity groups B, C and G in the glasshouse (see Results). Therefore it is likely that the sequences in the S6 group are also S-alleles.

In the region analysed, there is no evidence that the nine S-allele sequences obtained from self-compatible populations (seven from cirrhigerumGala and two from *Misopates orontium*) are pseudogenes. Intron splicing sites are conserved and there are no interruptions of the open reading frame. However, undetermined mutations elsewhere in the gene, including the promoter region, may have rendered this gene non-functional in these populations. The *Antirrhinum* and *Misopates* genera may be closely related. Thus there is little opportunity for these pseudogenes to accumulate neutral mutations relative to functional S-allele sequences.

7.4.2. *Genealogical analyses of Scrophulariaceae, Solanaceae and Rosaceae RNases*

With the exception of the *Petunia inflata* RNase X2 S-like RNase sequence (Richman *et al.* 1997), the other S-like RNase sequences of Solanaceae and Rosaceae are more similar to each other than to the S-alleles from different plant families, and therefore fall into a sister group of S-RNases (Xue *et al.* 1996, Richman *et al.* 1997; Ushijima *et al.* 1998). This has been taken as evidence that most S-like RNases may have diverged well

before the most recent common ancestor of the taxa from which they have been sampled. Although in *Antirrhinum* no RNases other than S-alleles have been yet described, a phylogenetic comparison among the putative S-alleles here described and other Solanaceae and Rosaceae S- and S- like RNases maybe informative, assuming that all *Antirrhinum* and *Misopates* S-like RNases originated before the split of Solanaceae / Scrophulariaceae and Rosaceae.

The relationship between the S-alleles representative of the most divergent lineages of *Antirrhinum,* Solanaceae and Rosaceae (Richman *et al.* 1997) and 14 S-like RNase sequences of different species of Solanaceae and Rosaceae (*Nicotiana tabacum*, *Nicotiana alata*, *Nicotiana glutinosa*, *Petunia inflata*, *Lycopersicum esculentum*, *Prunus dulcis*, *Malus domestica*, and *Pyrus pyrifolia*) is presented in Figure 7.3. All *Antirrhinum* sequences cluster together and define a sister group to all other sequences. That some S-like sequences are more similar to S-RNase sequences than to other S-like RNases suggests that not all S-like genes arose before the split of Solanaceae / Scrophulariaceae and Rosaceae. Therefore, although there is no definitive evidence that the *Antirrhinum* RNase sequences obtained here are S-like sequences, this conclusion is only tentative, since there are no *Antirrhinum* S-like sequences yet available.

7.4.3. Patterns of variability

Within groups sequences from unrelated individuals of the same species (i. e. from different populations) have very few differences. The nucleotide sequences are identical between graniticumB5 and graniticumS5; cirrhigerumGala9 and

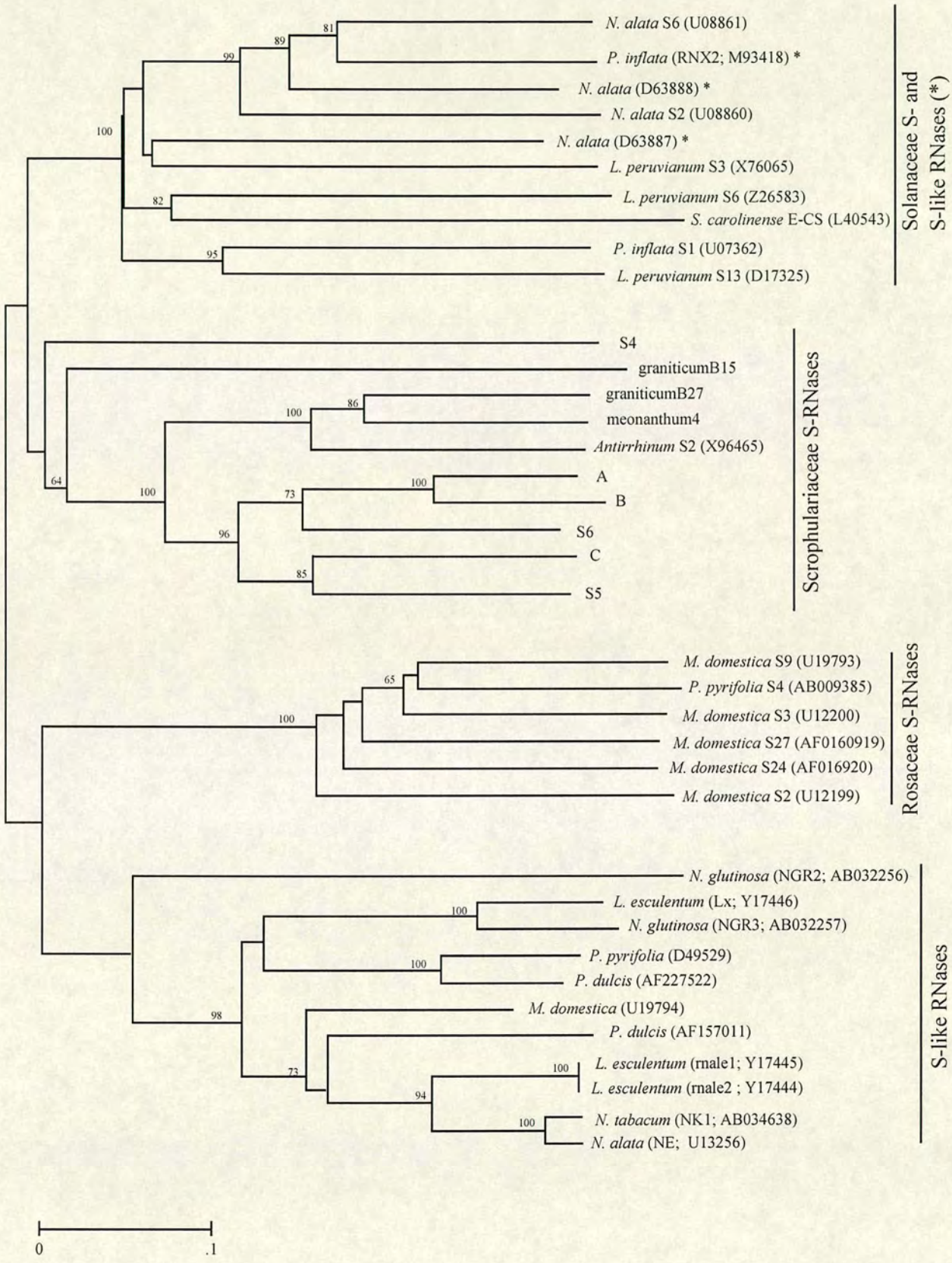


Fig. 7.3. Neighbor-joining tree using distances calculated as mean percentage amino acid differences, showing the relationships among Scrophulariaceae S-alleles, 13 S-allele sequences representative of the most divergent lineages of S-alleles from Solanaceae and Rosaceae (Richman *et al.* 1997) and 14 S-like RNase sequences of different species of Solanaceae and Rosaceae. Bootstrap replicates supporting the branches are shown for values greater than 62%.

cirrhigerumMuel7 differ by only one nucleotide position. If the sequences correspond to S-alleles with different specificities, this is to be expected, since the number of S-alleles with different specificities in natural populations is usually very large, even in small populations (see as examples Emerson 1938; Lewis 1948; Lawrence 1975). The effective population size of each allelic specificity is therefore small, and genetic drift will eliminate most variability within specificities (reviewed by Clark 1993). Within groups of sequences low levels of divergence are observed. *Antirrhinum* and *Misopates* species may therefore be a group of closely related species, as suggested by the analyses of other nuclear genes (Chapter 4).

Polymorphism levels vary among species (Table 7.7), but they are higher than observed for other *Antirrhinum* nuclear genes (Chapter 4, 5 and 6). However these are very similar to values obtained for other S-RNases in Solanaceae (Anderson *et al.* 1986; Clark *et al.* 1990; Kaufmann *et al.* 1991; Richman et Kohn 1999). There seems to be no association between polymorphism levels and the breeding system of the populations. For instance the self-incompatible *A. meonanthum* and *A. molle* populations have a lower estimated variability level than the largely self-compatible *A. majus* subsp. *cirrhigerum* (population code cirrhigerumGala). Sample sizes are however small and population structure may affect within species variability levels (Richman and Kohn 2000). Additional data on genes unlinked to the S-locus are needed in order to address this issue.

7.4.4. *S- allele sequence evolution*

As natural selection usually eliminates alleles coding for proteins with variant amino acid sequences, finding K_a/K_s ratios above one indicates that diversifying selection

Table 7.7. DNA sequence variability within *Antirrhinum* and *Misopates* species

Species	Population	Number of sequences	π_a (336.92)	π_s (92.88)	Number of groups	K_a (333.65)	K_s (92.35)
A. graniticum	graniticumB	9	0.3005	0.4793	6	0.3677	0.5659
A. meonanthum	meonanthum	4	0.0728	0.1384	3	0.0713	0.1387
A. majus susbsp. cirrhigerum	cirrhigerumGala	7	0.2670	0.4069	3	0.3502	0.5306
A. molle	molle	10	0.1515	0.2458	3	0.2460	0.4138
Misopates orontium	orontiumB	2	0.0089	0.0104	-	-	-

π_a is the number of nonsynonymous substitutions per nonsynonymous site; π_s is the the number of synonymous substitutions per synonymous site; K_a is the number of nonsynonymous substitutions per nonsynonymous site between sequences of different groups; K_s is the the number of synonymous substitutions per synonymous site between sequences of different groups; in brackets are the average number of sites analysed.

has occurred favouring new amino acid sequences. K_a and K_s values were calculated for all possible pairwise comparisons for the whole region analysed (Fig. 7.4A; the average K_a/K_s ratio is 0.460 ± 0.083), without conserved regions C1, C2 and C3, (Fig. 7.4B; the average K_a/K_s ratio is 0.534 ± 0.081) without amino acid sites that are conserved in all Scrophulariaceae S-allele sequences (Fig. 7.4C; the average K_a/K_s is 0.664 ± 0.130), or whithout both conserved and conservative substitutions (Fig. 7.4D; the average K_a/K_s is 0.748 ± 0.190). Although the average K_a/K_s value is always below 1, in both panel C and D (Fig. 7.4) some comparisons give values for this ratio over 1. Excess of non-synonymous changes have only been found for pairs of sequences for which the K_s value is above 52% (Fig. 7.4, C and D), and thus maybe due to saturation at silent sites. Furthermore in all analyses K_a and K_s values are highly significantly linearly correlated (R^2 varies from 0.81 to 0.93; $P < 0.001$ for all cases). A highly significant correlation and a slope of one is expected if amino replacement polymorphisms were neutral. When conserved and conservative amino acid sites are removed from the analyses (Fig 7.4D), the slope of the correlation is 0.71 ($R^2 = 0.81$). Therefore no strong evidence for diversifying selection is found, suggesting that only a fraction of all replacement polymorphisms are the target(s) of selection. Nevertheless the age of these putative S-alleles is only compatible with balancing selection acting on this locus. This is in contrast with Solanaceous species where K_a/K_s ratios over one are usually found (Richman *et al.* 1996b, Richman and Kohn 1999). Alleles that are more closely related tend to have higher ratios of non-synonymous to synonymous changes than do more distant alleles (Richman *et al.* 1996b; reviewed by Charlesworth and Guttman 1997). It is possible that

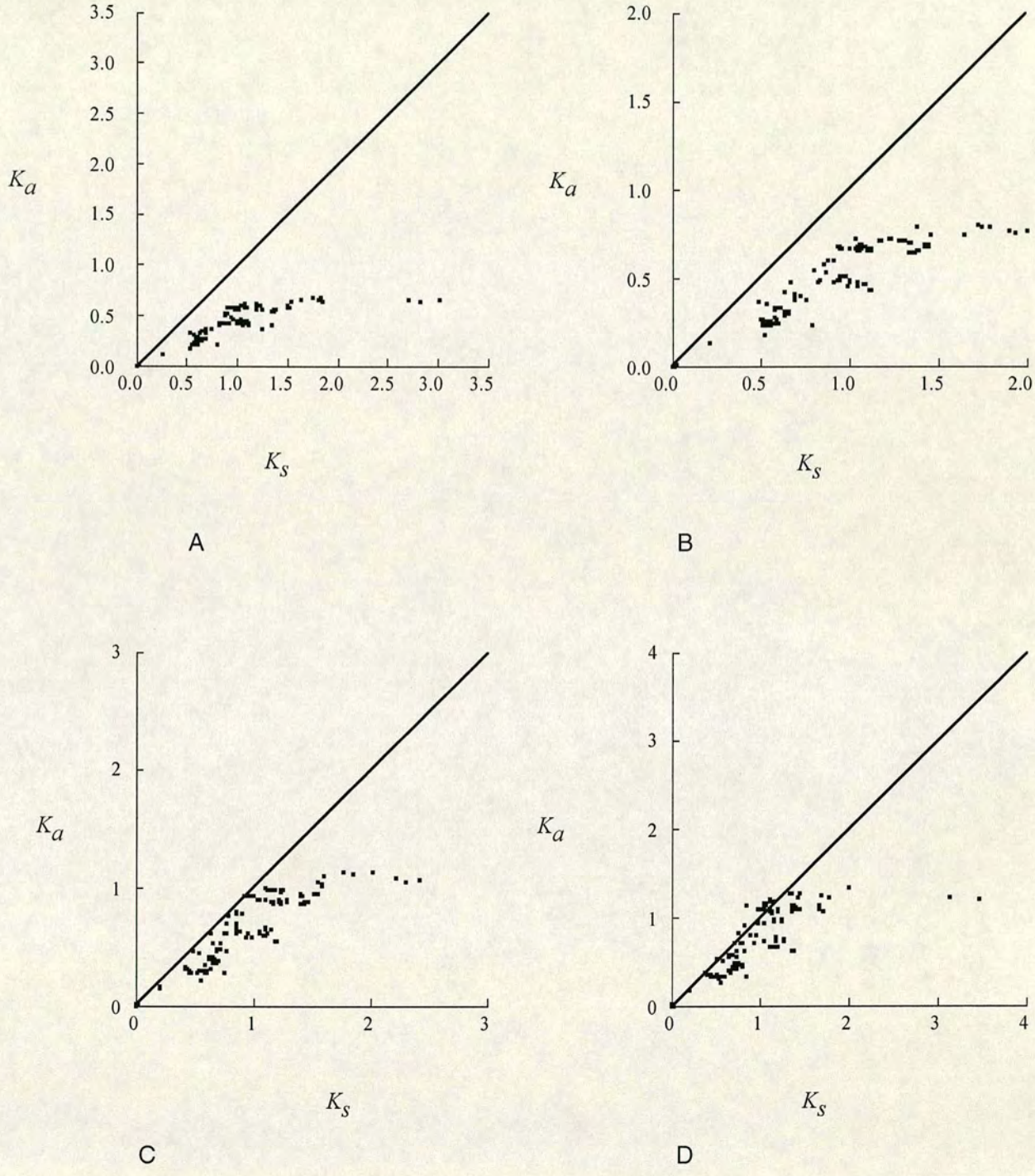


Fig. 7.4. K_a versus K_s values for pairwise sequence comparisons for *Antirrhinum* and *Misopates* S-allele sequences. (A) the whole region analysed. (B) excluding conserved regions C1, C2 and C3 (Xue *et al.* 1996). (C) excluding conserved amino acid replacements in Scrophulariaceae S-alleles. (D) excluding conserved positions and conservative amino acid replacements in Scrophulariaceae S-alleles.

once a new specificity type is established, the presence of conserved sites and sometimes restrictions in the amino acids allowed at variable sites limits further protein sequence divergence (Tanaka and Nei 1989). Therefore the average number of nonsynonymous differences between allele pairs increases more slowly than synonymous differences (Uyenoyama 1997). This is the opposite of the effect of saturation, and suggests diversifying selection, as differences accumulating neutrally should show a linear relation (Tanaka and Nei 1989).

Chapter 8

Conclusions

The large number of diploid closely related species with the same chromosome number (n= 8) in the *Antirrhinum* and *Misopates* genera, the short vegetative period, the small size of the plants, the fact that they are easy to cultivate, the fast germination, and the size and flower structure are some of the characteristics that make these genera suitable for genetical research (Chapter 2). Both genera have a Mediterranean range, most of the species living in the Iberian Peninsula that is the centre of diversification of the genera (Caputo *et al.* 1991), and large samples could be obtained.

Antirrhinum and *Misopates* species have different breeding systems, other factors that affect the levels and patterns of genetic diversity are not expected to vary greatly. Therefore studies within such genera should be ideal to test for effects of the mating system on levels and patterns of genetic diversity. However it is known from other taxa that individuals of different populations of the same species may sometimes have different breeding systems (Chapter 3). In *Antirrhinum majus* populations the breeding system varied from self-compatible to self-incompatible. Therefore these are ideal to

address the influence of the breeding system on levels and patterns of genetic diversity. Four floral features (corolla diameter and length; short and long anther-stigma separation) have been measured and their suitability as mating system indicators tested. None of these seem to be correlated with the breeding system.

In the first attempt to estimate the level of DNA diversity in the genus *Antirrhinum* and test for an effect of the breeding system on diversity, I have used members of the TCP gene family (Chapter 4). I present evidence that in *Antirrhinum* this gene family is composed of at least six members. Some of the genes have similar nucleotide sequences. This creates difficulties in obtaining data sets for variability studies. Nevertheless the few data obtained suggests that levels of nucleotide variability are low. Since the estimated nucleotide substitution rates for the TCP genes are also lower than most other estimates for plant nuclear genes (Gaut *et al.* 1996), it was then possible that members of this gene family were highly constrained at synonymous sites, or they may be slow evolving genes.

In my second attempt I used the *fil1* gene of *Antirrhinum* for which intron and 5' and 3' flanking coding sequence is available (Chapter 5). Therefore, regions with putatively different levels of constraint could be included in the diversity analyses. In addition, this gene has been reported to be a single copy gene. Detailed nucleotide diversity studies, however, revealed that the *fil1* gene of *Antirrhinum* is a member of a gene family composed of at least five genes. In four *Antirrhinum majus* populations with different mating systems, and one *A. graniticum* population, diversity is very low for the synonymous, intron and 3' flanking coding sequence. Divergence among *Antirrhinum* species and between *Antirrhinum* and *Digitalis* is also low. For three of these genes I also

obtained sequences from a more divergent member of the Scrophulariaceae, *Verbascum nigrum.* Compared with *Antirrhinum*, little divergence is again observed.

To assess the generality of low levels of genetic diversity and divergence I have extended these studies to further genes: *fil2*, *far, globosa* and *Adh* (Chapter 6). Detailed studies revealed that in *Antirrhinum* these genes belong to gene families. Low levels of divergence between *Antirrhinum* and *Verbascum* are observed for *fil2A*, *fil2B*, *far-L*, and *globosa1*. For *Adh* I could not determine orthology because repeated gene duplication and loss of elements in this gene family has occurred in the *Antirrhinum* and *Verbascum* lineages. Small *et al.* (2000) also found rapid copy number fluctuations for *Adh* in the genus *Gossypium*. The same pattern is also observed for other gene families (Clegg *et al.* 1997; Kramer *et al.* 1998; Meyers *et al.* 1999; Durbin *et al.* 2000; Lagercrantz and Axelsson 2000; Oberholzer *et al.* 2000; Pan *et al.* 2000; Theissen *et al.* 2000).

That nuclear genes are typically present in multiple copies in many plant genomes is not surprising, since the genome analysis of *Arabidopsis thaliana* revealed the presence of many multigene families (Delseny *et al.* 1997; Lin *et al.* 1999; Terryn *et al.* 1999). The GenBank analysis in Chapter 6 suggest that gene families are very common in Scrophulariaceae. In *Antirrhinum* very similar genes are present within gene families, in some cases differing in only one indel in the intron region, as in the *fil1* gene family (Chapter 4). The same pattern has been observed in other plant species as well. For instance, in *Mesembryanthemum crystallinum* (Aizoaceae), two of the six genes in the rubisco small subunit gene family, *rbc*S, were found to be identical at the nucleotide level, including an intron (DeRocher *et al.*, 1993). In *Arabidopsis thaliana* the BAC clone F9F13 contains 13 identical nucleotide sequences of a kinase gene (accession number

AL080253.1). Similar findings have been recently reported in humans (Eichler 1998; Horvath *et al.* 2000). This makes it difficult to establish orthology.

In general, there are no rules as to what level of difference suggests that more than a single gene is present. Length differences in introns often occur between different alleles at a single locus (Charlesworth *et al.* 1998; Liu *et al.* 1998; Small *et al.* 2000), so a length difference cannot be taken as evidence that a sequence is not allelic. Given the existence of diversity differences between loci under different selective regimes, and the fact that loci in genomic regions where recombination rarely occurs tend to show low diversity (Begun and Aquadro 1992; Stephan and Langley 1998), it is unlikely that there is any general way to recognise paralogy from sequence differences alone. Phylogenetic analyses, although generally useful, cannot reveal paralogy of very recent duplications unless many species are studied, as here. Additional data such as synteny, may help establish orthology (Doyle and Gaut, 2000). Even then, a thorough molecular characterisation of the different members of a gene family may also be required if the members of the gene family are arranged in tandem. Genetic evidence of segregation is not necessarily helpful, because, if there is genetic diversity at loci, one cannot always distinguish between a single locus and a tandem duplication. Southern analysis in most cases will fail to reveal the presence of more than one gene if sequence similarity is as high as in among some *Antirrhinum* gene families. To be rigorous, careful investigation of the presence or absence of sequences within and between individuals, i.e. of diversity, is therefore needed to establish whether a putative gene is truly a single locus. In most cases *Antirrhinum* genes have been isolated from cDNA libraries. A number of positive clones are usually isolated, and in some cases digested with several restriction enzymes,

but in general only a single clone is sequenced. Given the few differences between members of the gene families characterised here, these may be some of the reasons why the genes here studied were not previously recognised as belonging to gene families (Luo *et al.* 1996; Nacken *et al.* 1991; Tröbner *et al.* 1992; Steinmayr *et al.* 1994; Davies *et al.* 1999).

The low level of diversity and divergence observed for the *cyc*, *fil1; fil2; far* and *globosa* gene families also makes these gene families unsuitable for testing the effect of the breeding system on levels of variability. It is still unclear what is the cause of this pattern. Codon bias (reviewed by Sharp *et al.* 1993) and RNA structure (Parsch *et al.* 2000) were ruled out as the likely cause of the patterns observed (Chapter 5 and 6). Gene conversion between paralogous copies cannot explain either low diversity or divergence (Nagylaki and Petes 1982; Ohta 1981; 1984). An alternative resolution of the puzzle presented by these data is that these genes or species have an unusually low mutation rate. The genes for which the pattern of low diversity and divergence was observed are genes involved in floral development. Therefore this pattern could be a characteristic of families of developmental genes. Recently Purugganan (2000) suggested that regulatory genes are less variable than structural loci. This hypothesis could therefore be tested by analysing non developmental genes, especially those for which variability has been previously observed at the allozyme level. However, although there are currently 75 genes that have been partially or completely sequenced in *Antirrhinum* (GenBank release version 114.0, 08/09/00) the vast majority are genes involved in floral development. Many of the loci analysed in allozyme surveys have not yet been cloned and sequenced in any Scrophulariaceae species. *Adh* (a larger region than the 392 bp

obtained here; Chapter 6) may be suitable for testing for the effects of the mating system on levels and patterns of genetic diversity in *A. majus* populations. Copy number differences are unlikely to exist within species.

I have also obtained data on levels and patterns of variability at the gene responsible for self-incompatibility in *Antirrhinum*, the S-locus (de Nettancout 1977; Xue *et al.* 1996). Putative S-allele sequences were amplified from both self-incompatible and self-compatible populations and species. Although there is no evidence from the sequences of the region analysed that S-alleles from self-compatible populations are pseudogenes, they must be non-functional. The presence of similar S-allele sequences in *Antirrhinum* and *Misopates* suggests that these two genera are very closely related. The S-locus is highly variable, both at the nucleotide and amino acid level, compatible with balancing selection acting on this gene (reviewed by Clark 1993). The hypervariable regions of the S-locus have been suggested to be the target of selection (Richman *et al.* 1996b, Charlesworth and Guttman 1997, Richman and Kohn 1999). In Scrophulariaceae, analyses of K_a/ K_s ratios show no evidence for directional selection acting on these regions. However the two primer sets used here support the amplification of only about 10% of S-allele sequences. In the future, additional primer sets based on the conserved regions of Solanaceae and Rosaceae sequences can be used in order to amplify a larger variety of S-allele Scrophulariaceae sequences.

References

Abbott, R. J., and Gomes, M. F. 1988. Population genetic structure and outcrossing rate of *Arabidopsis thaliana* (L.) Heynh. Heredity 62:411-418.

Aceto, S., Caputo, P., Cozzolino, S., Gaudio, L., and Moretti, A. 1999. Phylogeny and evolution of *Orchis* and allied genera based on ITS DNA variation: morphological gaps and molecular continuity. Mol. Phylogenet. Evol. 13:67-76.

Aguadé, M., Miyashita, N., and Langley, C. H. 1989. Restriction-map variation at the *zeste-tko* region in natural populations of *Drosophila melanogaster*. Mol. Biol. Evol. 6:123-130.

Ai Y., Tsai D. S., Kao T. H. 1992. Cloning and sequencing of cDNAs encoding two S proteins of a self-compatible cultivar of *Petunia hybrida*. Plant Mol. Biol. 19:523-528.

Akashi, H., Kliman, M. R., and Eyre-Walker, A. 1998. Mutation pressure, natural selection, and evolution of base composition in *Drosophila*. Genetica 102/103:49-60

Alba, R., Kelmenson, P. M., Cordonnier-Pratt, M. M., and Pratt, L. H. 2000. The phytochrome gene family in tomato and the rapid differential evolution of this family in angiosperms. Mol. Biol. Evol. 17:362-373.

Almeida, J., Rocheta, M., and Galego, L. 1997. Genetic control of flower shape in *Antirrhinum majus*. Development 124:1387-1392.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25:3389-3402.

Amos, W., and Harwood, J. 1998. Factors affecting levels of genetic diversity in natural populations. Philos. Trans. R. Soc. Lond. B. Biol. Sci. 353:177-186.

Anderson, M. A., Hoggart, R. D., and Clarke, A. E. 1983. The possible role of lectins in mediating plant cell-cell interactions. In: Goldstein, I. J., and Etzler, M. E.

(eds). Chemical taxonomy, molecular biology and function of plant lectins, Alan R, Liss, New York pp. 143-161.

Anderson, M. A., Cornish, E. C., Mau, S. L. *et al.* 1986. Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in *Nicotiana alata*. Nature 321:38-44.

Anderson M. A., McFadden G. I., Bernatzky R. *et al.* 1989. Sequence variability of three alleles of the self-incompatibility gene of *Nicotiana alata*. Plant Cell 1:483:491.

Anonymous. 1983. Liste des plantes rares, menacées et endemiques en Europe. Coll. Sauvegarde de la Nature 27, Strasbourg.

Anonymous. 1992. Directive 92/43/CEE of the Council of the European Community on the Conservation of Habitats and Wild Fauna and Flora. European Community, Brussels.

Aquadro, C. F., and Begun, D. J. 1993. Evidence for and implications of genetic hitchhiking in the *Drosophila* genome. In: Takahata, N., and Clark, A. G. (eds). Mechanisms of molecular evolution. Sinauer, Sunderland, Mass, pp. 159-178.

Aquadro, C. F., Begun, D. J., and Kindahl, E. C. 1994. Selection, recombination and DNA polymorphism in *Drosophila*. In: Golding, B. (ed). Non-neutral evolution: theories and molecular data. Chapman and Hall, London, pp. 46-56.

Arasu, N. N. 1968. Self-incompatibility in angiosperms: a review. Genetica 39:1-24.

Arnheim, N. 1983. Concerted evolution of multigene families. In: Nei, M., and Koehn, R.K. (eds). Evolution of Genes and Proteins, Sinauer, Sunderland, Massachusetts, pp. 38-61.

Arroyo, M. T. K. 1973. Chiasma frequency evidence on the evolution of autogamy in *Limnanthes floccosa* (Limnanthaceae). Evolution 27:679-688.

Awadalla, P., and Ritland, K. 1997. Microsatellite variation and evolution in the *Mimulus guttatus* species complex with contrasting mating systems. Mol. Biol. Evol. 14:1023-1034.

Awadalla P., and Charlesworth D. 1999. Recombination and selection at *Brassica* self-incompatibility loci. Genetics 152: 413-425.

Bariola, P. A., Howard, C. J., Taylor, C. B., Verburg, M. T., Jaglan, V. D., Green, P. J. 1994. The *Arabidopsis* ribonuclease gene RNS1 is tightly controlled in response to phosphate limitation. Plant. J. 6:673-685.

Barreno, E., Bramwell, D., Cabezudo B., *et al.* 1984. Listado de las plantas endémicas, raras o amenazadas de España. Información Ambiental 3:369-377.

Barrett, S. C. H., and Shore, J. S. 1987. Variation and evolution of breeding systems in the *Turnera ulmifolia* L. complex (Turneraceae). Evolution 41:340-354.

Barrett, S. C. H., and Eckert, C. G. 1990. Variation and evolution of plant mating systems. In: Kawano, S. (ed.). Biological approaches and evolutionary trends in plants. Academic Press, New York, pp. 229-254.

Barrett, S. C. H., and Husband, B. C. 1990. Variation in outcrossing rates in *Eichhornia paniculata*: the role of demographic and reproductive factors. Pl. Species Biology 5:41-55.

Barrett, S. C. H., and Charlesworth, D. 1991. Effects of a change in the level of inbreeding on the genetic load. Nature 352:522-524.

Barton, N. H., and Whitlock, M. 1997. The evolution of metapopulations. In: Hanski, I. A., and Gilpin, M. E. (eds). Metapopulation biology. Academic Press, New York, pp. 183-210.

Bateman, A. J. 1952. Self-incompatibility systems in angiosperms. Heredity 6:285-310.

Bauer, V. L., and Aquadro, C. F. 1997. Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*. Mol. Biol. Evol. 14:1252-1257.

Baur, E. 1911. Einführung in die Experimentelle Verebungslehre. Berlin: Gebr. Borntraeger.

Baur, E. 1919. Über selbststerilität and über and kreuzungssterilität in der gattung *Antirrhinum*. Z. Indukt. Abstammungs-Vererbungsl. 21:48-52.

Baur, E. 1930. Mutations auslösung bei *Antirrhinum majus*. Zeitschrift für Botanik 23:676-702.

Begun, D. J., and Aquadro, C. F. 1992. Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. Nature 356:519-520.

Bonas, U., Sommer, H., Harrison, B. J., and Seadler, H. 1984. The transposable element TAM1 of *Antirrhinum majus* is 17 kb long. Mol. and General Genetics 194:138-143.

Boyes, D. C., Nasrallah, M. E., Vrebalov, J., and Nasrallah, J. B. 1997. The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. The Plant Cell 9:237-247.

Breese, E. L. 1959. Selection for differing degrees of outbreeding in *Nicotiana rustica*. Ann. Bot. 23:331-344.

Bremer, K. 2000. Early Cretaceous lineages of monocot flowering plants. Proc. Natl. Acad. Sci. USA 97:4707-4011.

Brewbaker, J. L. 1957. Pollen cytology and self-incompability systems in plants. J. Hered. 48:271-277.

Brewbaker, J. L. 1967. The distribution and phylogenetic significance of binucleate and trinucleate pollen grains in the angiosperms. Amer. J. Bot. 54:1069-1083.

Brieger, F. G. 1935. The inheritance of self-sterility and the peloric flower shape in *Antirrhinum*. Genetica 17:385-408.

Broothaerts W., Vanvinckenroye, P. Decock B., Van DammeJ., and Vendrig J. C. 1991. *Petunia hybrida* S-proteins: ribonuclease activity and the role of their glycan side chains in self-incompatibility. Sex. Plant Reprod. 4:258-266.

Broothaerts, W., Janssens, G. A., Proost, P., and Broekaert, W. F. 1995. cDNA cloning and molecular analysis of two self-incompatibility alleles from apple. Plant Mol. Biol. 27:499-511.

Brown, A. H. D. 1979. Enzyme polymorphism in plant populations. Theoretical Population Biology 15:1-42.

Brown, J. W., Smith, P., and Simpson, C. G. 1996. *Arabidopsis* consensus intron sequences. Plant Mol. Biol. 32:531-535.

Caputo, P., Cozzolino, S., and Campo, I. 1991-92. Relationships among the Italian species of genus *Antirrhinum* L. (Scrophulariaceae) based on chloroplast DNA restriction fragment lenght polymorphisms. Delpinoa 33-34:3-12.

Carpenter, R., and Coen, E. S. 1990. Floral homeotic mutations produced by transposon-mutagenesis in *Antirrhinum majus*. Genes Dev. 4:1483-1493.

Charlesworth, D. 1985. Distribution of dioecy and self-incompatibility in angiosperms. In: Greenwood, J. J., and Slatkin, M. (eds). Evolution-essays in honour of John Maynard Smith. Cambridge Univ. Press, Cambridge, pp. 237-268.

Charlesworth, D., and Charlesworth, B. 1987. Inbreeding depression and its evolutionary consequences. Annu. Rev. Ecol. Syst. 18:237-268.

Charlesworth, D., and Guttman, D. S. 1997. Seeing selection in S allele sequences. Curr. Biol. 7:R34-37.

Charlesworth, D., and Charlesworth, B. 1995. Transposable elements in inbreeding and outbreeding populations. Genetics 140:415-417.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303.

Charlesworth, B., Charlesworth, D., and Morgan, M. T. 1995. The pattern of neutral molecular variation under the background selection model. Genetics 141:1619-1632.

Charleswoth D., and Pannell J. R. 2000. Mating systems and population genetics structure. Submitted.

Charlesworth, B., Nordborg, M., and Charlesworth, D. 1997. The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. Genetical Research 70:155-174.

Charlesworth, D., and Charlesworth, B. 1998. Sequence variation: looking for effects of genetic linkage. Current Biology 8:R658-661.

Charlesworth, D., and Awadalla, P. 1998. The molecular population genetics of flowering plant self-incompatibility polymorphisms. Heredity 81:1-9.

Charlesworth, D., and Yang, Z. 1998. Allozyme diversity in *Leavenworthia* populations with different inbreeding levels. Heredity 81:453-461.

Charlesworth, D., Liu, F. L., and Zhang, L. 1998. The evolution of the alcohol dehydrogenase gene family by loss of introns in plants of the genus *Leavenworthia* (*Brassicaceae*). Mol. Biol. Evol. 15:552-559.

Chase, M. W., Soltis, D. E., Olmstead, R. G., *et al.* 1993. Phylogenetics of seed plants: an analysis of nucleotide sequences from the plastid gene *rbc*L. Ann. Missouri Bot. Gard. 80:528-580.

Chiapello, H., Lisacek, F., Caboche, M., and Henaut, A. 1998. Codon usage and gene function are related in sequences of *Arabidopsis thaliana*. Gene 209:GC1-38.

Citerne, H. 1999. Examination of *cycloidea*-like genes in Gesneriaceae and their implication in the control of zygomorphy. MSc. Thesis. University of Edinburgh.

Citerne, H. L., Moller, M., and Cronk, Q. C. B. 2000. Diversity of *cycloidea*-like genes in Gesneriaceae in relation to floral symmetry. Annals of Botany 86:167-176.

Clark, R. E., Okuley, J. J., Collins, P. D., and Sims, T. L. 1990. Sequence variability and developmental expression of S-alleles in self-incompatible and pseudo-self-incompatible *Petunia*. Plant Cell 2:815-826.

Clark, A. G. 1993. Evolutionary inferences from molecular characterization of self-incompatibility alleles. In: Takahata, N., and Clark, A. G. (eds.). Mechanisms of molecular evolution. Sinauer Associates, Sunderland, pp. 79-108.

Clark, A. G. 1996. Population genetic aspects of gametophytic self-incompatibility. Plant Species Biol. 11:13-21.

Clark, A. G. 1997. Neutral behavior of shared polymorphism. Proc. Natl. Acad. Sci. USA 94:7730-7734.

Clegg, M. T. 1997. Plant genetic diversity and the struggle to measure selection. J. Hered. 88:1-7.

Clegg, M.T., Cummings, M. P., and Durbin, M.L. 1997. The evolution of plant nuclear genes. Proc. Natl. Acad. Sci. USA 94:7791-7798.

Coen, E. S., and Carpenter, R. 1986. Transposable elements in *Antirrhinum majus*: generators of genetic diversity. Trends in Genetics 2:292-296.

Coleman, C. A., and Kao, T. H. 1992. The flanking regions of *Petunia inflata* S alleles are heterogeneous and contain repetitive sequences. Pl. Molec. Biol. 18:725-737.

Comeron, J. M., Kreitman, M., and Aguadé, M. 1999. Natural selection on synonymous sites is correlated with gene length and recombination in *Drosophila*. Genetics 151:239-249.

Cornish, E. C., Pettitt, J. M., Bonig, I., and Clark, A. E. 1987. Developmentally controlled expression of a gene associated with self-incompatibility in *Nicotiana alata*. Nature 326:99-102.

Crane, P. R., Friis, E. M., and Pedersen, K. R. 1995. The origin and early diversification of angiosperms. Nature 374:27-33.

Cubas, P., Lauter, N., Doebley, J., and Coen, E. 1999a. The TCP domain: a motif in proteins regulating plant growth and development. The Plant J. 18:215-222.

Cubas, P., Vincent, C., and Coen, E. 1999b. An epigenetic mutation responsible for natural variation in floral symmetry. Nature 401:157-161.

Cummings, M. P., and Clegg, M. T. 1998. Nucleotide sequence diversity at the alcohol dehydrogenase 1 locus in wild barley (*Hordeum vulgare* ssp. *spontaneum*): an evaluation of the background selection hypothesis. Proc. Natl. Acad. Sci. USA 95:5637-5642.

Davies, B., Motte, P., Keck, E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. 1999. PLENA and FARINELLI: redundancy and regulatory interactions between two *Antirrhinum* MADS-box factors controlling flower development. EMBO J. 18:4023-4034.

Dayhoff, M. O., Schwartz, R. M., and Orcutt, B. C. 1979. A model of evolutionary change in proteins. In: Dayhoff, M. O. (ed) Atlas of protein sequence and

structure vol 5 (suppl 3). National Biomedical Research Foundation, Washington, DC, pp. 345-352.

Delavault, P., Estabrook, E., Albrecht, H., Wrobel, R., and Yoder J. I. 1998. Host-root exudates increase gene expression of asparagine synthetase in the roots of a hemiparasitic plant *Triphysaria versicolor* (Scrophulariaceae). Gene 222:155-162.

Delseny, M., Cooke, R., Raynall, M., and Grellet, F. 1997. The *Arabidopsis thaliana* cDNA sequencing projects. FEBS Lett. 405:129-132.

de Nettancourt, D. 1977. Incompatibility in Angiosperms. Springer-Verlag. Berlin.

DeRocher, E. J., Quigley, F., Mache, R., and Bohnert, H. J. 1993. The six genes of the Rubisco small subunit multigene family from *Mesembryanthemum crystallinum*, a facultative CAM plant. Mol. Gen. Genet. 239:450-462.

Doaigey, A. R., and Harkiss, K. J. 1991. Application of epidermal characters to the taxonomy of European species of *Antirrhinum* (Schrophulariaceae). Nord. J. Bot. 11:513-524.

Dodds, P. N., Clarke, A. E., and Newbigin, E. 1996. Molecular characterisation of an S-like RNase of *Nicotiana alata* that is induced by phosphate starvation. Plant Mol. Biol. 31:227-238.

Doebley, J., Stec, A., and Hubbard, L. 1997. The evolution of apical dominance in maize. Nature 386:485-488.

Doyle, J. A. 1978. Origin of angiosperms. Annu. Rev. Ecol. Syst. 9:365-392.

Doyle, J. J., and Doyle, J. L. 1999. Nuclear protein-coding genes in phylogeny reconstruction and homology assessment: some examples from Leguminosae. In: Hollingsworth, P. M., Bateman, R. J., and Gornall, R. J. (eds). Molecular systematics and plant evolution. Taylor and Francis, London, pp. 229-254.

Doyle, J. J., and Gaut, B. S. 2000. Evolution of genes and taxa: a primer. Plant Mol. Biol. 42:1-23.

Durbin, M. L., McCaig, B., and Clegg, M. T. 2000. Molecular evolution of the chalcone synthase multigene family in the morning glory genome. Plant Mol. Biol. 42:79-92.

Duret, L., and Mouchiroud, D. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. Proc. Natl. Acad. Sci. USA 96:4482-4487.

Dvorak, J., Luo, M. C., and Yang, Z. L. 1998. Restriction fragment length polymorphism and divergence in the genomic regions of high and low recombination in self-fertilizing and cross-fertilizing *Aegilops* species. Genetics 148:423-434.

East, E. M., and Mangelsdorf, A. J. 1925. A new interpretation of the hereditary behaviour of self-sterile plants. Proc. Nat. Acad. Sci. 11:166-171.

East, E. M. 1940. The distribution of self-sterility in the flowering plants. Proc. Amer. Philos. Soc. 82:449-518.

Ebert, P. R., Anderson, M. A., Bernatzky, R., Altschuler, M., and Clarke, A. E. 1989. Genetic polymorphism of self-incompatibility in flowering plants. Cell. 56:255-262.

Eichler, E. E. 1998. Masquerading repeats: paralogous pitfalls of the human genome. Genome Res. 8:758-762.

Elisens, W. J., and Crawford, D. J. 1988. Genetic variation and differentiation in the genus *Mabrya* (Scrophulariaceae-Antirrhineae): systematic and evolutionary inferences. Am. J. Bot. 75:85-96.

Elisens, W. J. 1992. Genetic divergence in *Galvezia* (Scrophulariaceae): evolutionary and biogeographic relationships among South American and Galapagos species. Am. J. Bot. 79:198-206.

Elisens, W. J., and Nelson, A. D. 1993. Morphological and isozyme divergence in *Gambelia* (Scrophulariaceae): species delimitation and relationships. Syst. Bot. 18:454-468.

Emerson, S. 1938. The genetics of self-incompatibility in *Oenothera organensis*. Genetics 23:201-202.

Estoup, A., and Cornuet, J. M. 1999. Microsatellite evolution: inferences from population data. In: Goldstein, D. B., and Schlötterer, C. (eds). Microsatellites. Evolution and applications. Oxford University Press, Oxford, pp. 50-65.

Fenster, C. B., and Ritland, K. 1992. Chloroplast DNA and isozyme diversity in two *Mimulus* species (Scrophulariaceae) with contrasting mating systems. Am. J. Bot. 79:1440-1447.

Fernandes, A., Queiros, M., and Santos, M. F. 1977. Contribution to the cyto taxonomic knowledge of spermatophyta of Portugal . XV Scrophulariaceae. Boletin da Sociedade Broteriana 51:37-90.

Filatov, D. A., and Charlesworth, D. 1999. DNA polymorphism, haplotype structure and balancing selection in the *Leavenworthia PgiC* locus. Genetics 153:1423-1434.

Florea, C. S., and Timko, M. P. 1997. Actin genes with unusual organization in the parasitic angiosperm *Striga asiatica* L. (Kuntze). Gene 186:127-133.

Franco, J. A. 1971. Nova Flora de Portugal (Continente e Açores): 1. Asa Editores, Lisboa.

Franklin-Tong, V. E. and Franklin, F. C. H. 1993. Gametophytic self-incompatibility: contrasting mechanisms for *Nicotiana* and *Papaver*. Trends Cell Biol. 3:340-345.

Fu, Y. X., and Li, W. H. 1993. Statistical tests of neutrality of mutations. Genetics 133:693-709.

Fukada-Tanaka, S., Hoshino, A., Hisatomi, Y., Habu, Y., Hasebe, M., and Iida, S. 1997. Identification of new chalcone synthase genes for flower pigmentation in the Japanese and common morning glories. Plant Cell Physiol. 38:754-758.

Futamura, N., Mori, H., Kouchi, H., and Shinohara, K. 2000. Male flower-specific expression of genes for polygalacturonase, pectin methylesterase and beta-1,3

glucanase in a dioecious willow (*Salix gilgiana* Seemen). Plant Cell Physiol. 41:16-26.

Galloway, G. L., Malmberg, R. L., and Price, R. A. 1998. Phylogenetic utility of the nuclear gene Arginine decarboxylase: an example from Brassicaceae. Mol. Biol. and Evol. 15:1312-1320.

Gaut, B. S., Morton, B. R., Mccaig, B. C., and Clegg, M. T. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parellel rate differnces at the plastid gene *rbc*L. Proc. Nat. Acad. Sci. USA 93:10274-10279.

Gaut, B. S. 1998. Molecular clocks and nucleotide sustitution rates in higher plants. Evol. Biol. 30: 93-120.

Gaut, B. S., Peek, A. S., Morton, B. R., and Clegg, M.T. 1999. Patterns of genetic diversification within the *Adh* gene family in the grasses (Poaceae). Mol. Biol. Evol. 16:1086-1097.

Gianfagna, T., Janes, H. W., and Chen, B. Y. 1998. PCR cloning and characterization of multiple ADP-glucose pyrophosphorylase cDNAs from tomato. Plant Science136:59-67.

Gibbs, P. E. 1988. Self-incompatibility mechanisms in flowering plants: some complications and clarifications. Lagascalia 15:17-28.

Gilbert, D. G. 1995. Seqpup version 0.6f: a biosequence editor and analysis application. http://sunsite.sut.ac.jp/pub/academic/biology/molbio/seqp up/

Glover, D. E., and Barrett, S. C. H. 1986. Variation in the mating system of *Eichhorniapaniculata* (Spreng.) Solms. Pontederiaceae. Evolution 40:1122-1131.

Godron, D. A. 1863. Des hybrides vegetaux considérés au point de vue de la fécondité et de la perpétuité ou non-prepétuite de leurs charactères. Ann. Sci. Nat. IV, Série Bot. Toms XIX:135-179.

Golz, J. F., Clarke, A. E., and Newbigin, E. 1995. Self-incompatibility in flowering plants. Curr. Opin. Genet. Dev. 5:640-645.

Gong, Z., Yamazaki, M., Sugiyama, M., Tanaka, Y., and Saito, K. 1997. Cloning and molecular analysis of structural genes involved in anthocyanin biosynthesis and expressed in a forma-specific manner in *Perilla frutescens*. Plant Mol. Biol. 35:915-927.

Goremykin, V. V., Hansmann, S., and Martin, W. F. 1997. Evolutionary analysis of 58 proteins encoded in six completely sequenced chloroplast genomes: revised molecular estimates of two seed plant divergence times. Plant Syst. and Evol. 206:337-351.

Grant, D., Cregan, P., and Shoemaker, R. C. 2000. Genome organization in dicots: Genome duplication in *Arabidopsis* and synteny between soybean and *Arabidopsis*. Proc. Nat. Acad. Sci. USA 97:4168-4173.

Green P. J. 1994. The ribonucleases of higher plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. 45:421-445.

Güemes, J., Mateu, I., and Sánchez, P. 1993. *Antirrhinum subbaeticum* Güemes, Mateu and Sánchez (Scrophulariaceae), especie nueva de la Peninsula Ibérica. Anales de Jardin Botánico de Madrid 51:237-247.

Hadfield, K. A., and Bennett, A. B. 1998. Polygalacturonases: many genes in search of a function. Plant Physiol. 117:337-343.

Halliday, G., and Beadle, M. 1983. Flora Europaea. Cambridge University Press, Cambridge.

Hamrick, J. L., and Godt, M. J. 1990. Allozyme diversity in plant species. In: Brown, A. H. D., Clegg, M. T., Kahler, A. L., and Weir, B. S. (eds). Plant population genetics, breeding and genetic resources. Sinauer, Sunderland, Mass, pp. 43-63.

Hamrick, J. L., and Godt, M. J. W. 1996. Effects of life history traits on genetic diversity in plant species. Phil. Trans. Roy. Soc. Lond. B. Biol. Sci. 351:1291-1298.

Hayman, D. L. 1956. The genetical control of incompability in *Phalaris coerulescens*. Desf. Aust. J. Biol. Sci. 9:321-331.

Hayman D. L., and Richter, J. 1992. Mutation affecting self-incompatibility in *Phalaris coerulescens* Desf. Heredity 68:495-503.

Harrison, B. J., and Darby, L. A. 1955. Uninateral hybridization. Nature 176:982.

Harte, C. 1974. *Antirrhinum majus*. In: King, R. C. (ed.). Handbook of genetics. Vol.2. Publishing corporation, New York.

Heywood, V. H. 1985. Flowering Plants of the World. Croom Helm Publishers, England.

Herrmann, V. H. 1973. Selbst-und kreuzungsinkompatibilitat verschiedener *Antirrhinum*- arten. Biol. Zentralbl. 92:773-777.

Hickey, L. J., and Doyle, J. A. 1977. Early Cretaceous fossil evidence for angiosperm evolution. Bot. Review 43:3-104.

Hinata, K., Watanabe, M., Yamakawa, S., Satta, Y., and Isogai, A. 1995. Evolutionary aspects of the S-related genes of the *Brassica* self-incompatibility system: synonymous and nonsynonymous base substitutions. Genetics 140:1099-1104.

Hodges, S. A., and Arnold, M. L. 1994. Columbines: a geographically widespread species flock. Proc. Natl. Acad. Sci. USA 91:5129-5132.

Hodges, S. A., and Arnold, M. L. 1995. Spurring plant diversification: are floral nectar spurs a key innovation ? Proc. R. Soc. Lond. B. 262:343-348.

Holtsford, T. P., and Ellstrand, N. C. 1989. Genetic and environmental variation in floral traits affecting outcrossing rate in *Clarkia tembloriensis* (Onagraceae). Evolution 46:216-225.

Horvath, J. E., Schwartz, S., and Eichler, E E. 2000. The mosaic structure of human pericentromeric DNA: a strategy for characterizing complex regions of the human genome. Genome Res. 10:839-852.

Huang, S., Lee, H. S., Karunanandaam B., and Kao, T. H. 1994. Ribonuclease activity of *Petunia inflata* S proteins is essential for rejection of self-pollen. Plant Cell. 6:1021-1028.

Hudson, R. R. 1983. Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. 23:183-201.

Hudson, R. R., Kreitman, M., and Aguadé, M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.

Hudson, R. R. 1990. Gene genealogies and the coalescent process. Oxford Surv. Evol. Biol. 7:1-45.

Hudson, R. R., and Kaplan, N. L. 1994. Gene trees with background selection. In: Golding, B. (ed.). Non-neutral evolution: theories and molecular data. Chapman and Hall, New York.

Hudson, R. R., Bailey, K., Skarecky, D., Kwiatowski, J., and Ayala, F. J. 1994. Evidence for positive selection in the superoxide dismutase (*Sod*) region of *Drosophila melanogaster*. Genetics 136:1329-1340.

Hudson, R. R., and Kaplan, N. L. 1995. Deleterious background selection with recombination. Genetics 141:1605-1617.

Hughes, A., and Nei, M. 1989. Nucleotide substitution at major histocompatibility complex class II loci: evidence for over-dominant selection. Proc. Natl. Acad. Sci. USA 86:958-962.

Hughes, A., Ota, T., and Nei, M. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. 76:515-524.

Ingram, G. C., Doyle, S., Carpenter, R., Schultz, E. A., Simon, R., and Coen, E.S. 1997. Dual role for *fimbriata* in regulating floral homeotic genes and cell division in *Antirrhinum*. EMBO J. 16:6521-6534.

Ioerger, T. R., Clark, A. G., and Kao, T. H. 1990. Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. Proc. Natl. Acad. Sci. U S A 87:9732-9735.

Ioerger, T. R., Gohlke, J. R., and Kao, T-H. 1991. Primary structural features of the self-incompatibility protein in Solanaceae. Sex. Plant Reprod. 4:81-87.

Iturriaga, G., Leyns, L., Villegas, A., Gharaibeh, R., Salamini, F., and Bartels, D. 1996. A family of novel *myb*-related genes from the resurrection plant *Craterostigma*

plantagineum are specifically expressed in callus and roots in response to ABA or desiccation. Plant Mol. Biol. 32:707-716.

Jahnen, W., Batterham, M. P., Clarke, A. E., Moritz, R. L., and Simpson, R. J. 1989. Identification, isolation, and N-terminal sequencing of style glycoproteins associated with self-incompatibility in *Nicotiana alata*. Plant Cell. 1:493-499.

Jost, W., Bak, H., Glund, K., Terpstra, P., Beintema, J. J. 1991. Amino acid sequence of an extracellular, phosphate-starvation-induced ribonuclease from cultured tomato (*Lycopersicon esculentum*) cells. Eur. J. Biochem. 198:1-6.

Jukes, T. H., and Cantor, C. R. 1969. Evolution of protein molecules. In: Munro, H. N. (ed.), Mammalian protein metabolism, Academic Press, New York, pp. 21.

Kamboj, R. K., and Jackson, J. F. 1986. Self-incompatibility alleles control a low molecular weight, basic protein in pistils of *Petunia hybrida*. Theor. Appl. Genet. 71:815-819.

Kampny, C. M. 1995. Pollination and flower diversity in Scrophulariaceae. The Botanical Review 61:350-366.

Kanazin, V., Marek, L. F., and Shoemaker, R. C. 1996. Resistance gene analogs are conserved and clustered in soybean. Proc. Natl. Acad. Sci. USA 93:11746-11750.

Kao, T-H, and McCubbin, A. G. 1996. How flowering plants discriminate between self and non-self pollen to prevent inbreeding. Proc. Natl. Acad. Sci. U S A 93:12059-12065.

Kao T. H., and McCubbin A. G. 1997. Molecular and biochemical bases of gametophytic self-incompatibility in Solanaceae. Plant Physiol. Biochem. 35:171-176.

Karunanandaa, B., Huang, S., and Kao T. H. 1994. Carbohydrate moiety of the *Petunia inflata* S3 protein is not required for self-incompatibility interactions between pollen and pistil. Plant Cell. 6:1933-1940.

Kaufmann, H., Salamini, F., and Thompson, R. D. 1991. Sequence variability and gene structure at the self-incompatibility locus of *Solanum tuberosum*. Mol. Gen. Genet. 226:457-466.

Kawata, Y., Sakiyama, F, Hayashi, F., Kyogoku, Y. 1990. Indentification of two essential histidine residues of ribonuclease T2 from *Aspergillus oryzae*. Eur. J. Biochem. 187:255-262.

Kelly, J. K. 1997. A test of neutrality based on interlocus associations. Genetics 146: 1197-1206.

Kelly, A. J., and Willis, J. H. 1998. Polymorphic microsatellite loci in *Mimulus guttatus* and related species. Molecular Ecology 7:769-774.

Kheyr-Pour, A., Bintrim, S. B., Ioeger, T. R., Remy, R., Hammond, S. A., and Kao, T. H. 1990. Sequence diversity of pistil S-proteins associated with gametophytic self-incompatibility in *Nicotiana alata*. Sex. Plant Reprod. 3:88-97.

Kimura, M. 1983. The Neutral Theory of Molecular Evolution. Cambridge University Press, Cambridge.

Kimura, M., and Ohta, T. 1971. Theoretical topics in population genetics. Princeton University Press, Princeton, New York.

Kliman, R. M., and Hey, J. 1993. Reduced natural selection associated with low recombination in *Drosophila melanogaster*. Mol. Biol. Evol. 10:1239-1258.

Kosugi, S., and Ohashi, Y. 1997. PCF1 and PCF2 specifically bind to *cis* elements in the rice proliferating cell nuclear antigen gene. Plant Cell 9:1607-1619.

Kovaleva, L. V., and Musatova, N. A. 1975. Ribonuclease activity in the generative organs of Petunia in connection with the formation of the mechanism of autoincompatibility. Dokl. Nauk. Akad. SSSR 223:762.

Kovaleva, L. V., Milyaeva, E. L., and Chailakhyan, M. K. H. 1978. Overcoming self-incompatibility by inhibitors of nucleic acid and protein metabolism. Phytomorphology 28: 445.

Kovaleva, L. V. 1983. The protein self-incompatibility factor in petunia styles. Dokl. Nauk. Akad. SSSR 272:1017.

Kraft, T., Sall, T., Magnusson-Rading, I., Nilsson, N. O., and Hallden, C. 1998. Positive correlation between recombination rates and levels of genetic variation in natural populations of sea beet (*Beta vulgaris* subsp. *maritima*). Genetics 150:1239-1244

Kramer, E. M., Dorit, R. L., and Irish, V. F. 1998. Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the *APETALA*3 and *PISTILLATA* MADS-box gene lineages. Genetics 149:765-783.

Kreitam, M., and Akashi, H. 1995. Molecular evidence for natural selection. Ann. Rev. Ecol. Syst. 26:403-422.

Kreitman, M., and Hudson, R. R. 1991. Inferring the evolutionary histories of the *Adh* and *Adh*-dup loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics 127:565-582.

Kumar, S., Tamura, K., and Nei, M. 1994. MEGA- molecular evolutionary genetics analysis software for microcomputers. Computer Applications in the Biosciences 10: 189-191.

Kuroda, S., Norioka, S., Mitta, M., Kato, I., and Sakiyama, F. 1994. Primary structure of a novel stylar Rnase unassociated with self-incompatibility in tobacco plant, *Nicotiana alata*. J. Protein Chem. 13:438-439.

Lagercrantz, U., and Axelsson, T. 2000. Rapid evolution of the family of CONSTANS LIKE genes in plants. Mol. Biol. Evol. 17:1499-1507.

Laguna, E. 1994. Flora vascular rara, endémica o amenazada de la Comunidad Valenciana.GeneralitatValenciana.

Lavin, M., Eshbaugh, E., Hu, J., Mathews, S., and Sharrock, R. A. 1998. Monophyletic subgroups of the tribe Millettieae (Leguminosae) as revealed by phytochrome nucleotide sequence data. Am. J. Bot. 85:412-433.

Lawrence, M. J. 1975. The genetics of self-incompatibility in *Papaver rhoeas*. Proc. Roy. Soc. London B. 188:275-285.

Lawrence, M. J. 1984. The genetical analysis of ecological traits. In: Shorrocks, B. (ed.). Evolutionary ecology. Blackwell, Oxford, pp. 27-63.

Lee, H. S., Singh, A., and Kao, T-H. 1992. RNase X2, a pistil-specific ribonuclease from *Petunia inflata*, shares sequence similarity with solanaceous S proteins. Plant Mol. Biol. 20:1131-1141.

Lee, Y. H., and Vaquier, V. D. 1992. The divergence of species-specific abalone sperm lysins is promoted by positive Darwinian selection. Biol. Bull. 182:97-104.

Lers, A., Khalchitski, A., Lomaniec, E., Burd, S., and Green, P. J. 1998. Senescence-induced RNases in tomato. Plant Mol. Biol. 36:439-449.

Lewis, D. 1948. Mutation of the incompatibility gene. I. Spontaneous mutation rate. Heredity 2:219-236.

Lewis, D. 1954. Comparative incompatibility in angiosperms and fungi. Advanc. Genet. 6:235-285.

Lewontin, R. C. 1964. The interaction of selection and linkage. I. General considerations: heterotic models. Genetics 49:49-67.

Li, W. H. 1997. Molecular Evolution. Sinauer Associates, Sunderland, MA.

Lin, X., Kaul, S., Rounsley, S., *et al.* 1999. Sequence and analysis of chromosome 2 of the plant *Arabidopsis thaliana*. Nature 402:761-768.

Lis, H., and Sharon, N. 1993. Protein glycosylation. Structural and functional aspects. Eur. J. Biochem. 218:1-27.

Liu, F-L., Zhang, L., and Charlesworth, D. 1998. Genetic diversity in *Leavenworthia* populations with different inbreeding levels. Proc. R. Soc. Lond. B. Biol. Sci. 22:293-301.

Liu, F-L., Charlesworth, D., and Kreitman, M. 1999. Joint influence of mating system and selection on nucleotide diversity at the phosphoglucose isomerase locus in the plant genus *Leavenworthia*. Genetics 151:343-357.

Lloyd, D. G. 1965. Evolution of self-compatibility and racial differentiation in *Leavenworthia* (Cruciferae). Am. J. Bot. 78:270-287.

Löffler, A., Abel, S. Jost, W., Beintema, J. J., and Glund K. 1992. Phosphate-regulated induction of intacellular ribonucleases in cultured tomato (*Lycopersicon esculentum*) cells. Plant Physiol. 98:1472-1478.

Löffler, A., Glund, K., and Irie, M. 1993. Amino acid sequence of an intracellular, phosphate-starvation-induced ribonuclease from cultured tomato (*Lycopersicon esculentum*) cells. Eur. J. Biochem. 214:627-633.

Lotsy, J. P. 1911. Hybrides entre espèces d'A*ntirrhinum*. In IVth Conférence Internationale de Génétique, Paris.

Lundqvist, A. 1975. Complex self-incompability systems in angiosperms. Proc. R. Soc. London B. 188:235-245.

Luo, D., Carpenter, R., Vincent, C., Copsey, L., and Coen, E. 1996. Origin of floral asymmetry in *Antirrhinum*. Nature 383:794-799.

Luo, D., Carpenter, R., Copsey, L., Vincent, C., Clark, J., and Coen, E. 1999. Control of organ asymmetry in flowers of *Antirrhinum*. Cell 99:367-376.

Ma, R. C., and Oliveira, M. M. 2000. The RNase PD2 gene of almond (*Prunus dulcis*) represents an evolutionarily distinct class of S-like RNase genes. Mol. Gen. Genet. 263:925-933.

MacDonald, P. M. 1990. *bicoid* mRNA localization signal: phylogenetic conservation of function and RNA secondary structure. Development 110:161-171.

MacDonald, P. M., and Struhl, G. 1988. *Cis*-acting sequences responsible for anterior localization of *bicoid* mRNA in *Drosophila* embryos. Nature 336:595-598.

Mahal, C., and Pal, M. 1984. *Antirrhinum orontium* complex: biosystematic studies. Proc. Indian Acad. Sci. 93:455-458.

McDonald, J.H., and Kreitman, M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. Nature 351:652-654.

Martin, C., Carpenter, R., Sommer, H., Saedler, H., and Coen, E. S. 1985. Molecular analysis of instability in flower pigmentation in *Antirrhinum majus*, following isolation of the *pallida* locus by transposon tagging. EMBO J. 4:1625-1630.

Martin, C., Prescott, A., Mackay, S., Bartlett, J., and Vrijlandt, E. 1991. The control of anthocyanin biosynthesis in flowers of *Antirrhinum majus*. Plant J. 1:37-49.

Mateu-Andres, I. 1999. Allozymic variation and divergence in three species of *Antirrhinum* L. (Scrophulariaceae-Antirrhineae). Bot. J. Linnean. Soc. 131:187-199.

Mather, K. 1947. Species crosses in *Antirrhinum*. I. Genetic isolation of the species *Majus*, *Glutinosum*, and *Orontium*. Heredity 1:175-186.

Matton, D. P., Nass, N., Clark, A. E., and Newbigin, E. 1994. Self-incompatibility: how plants avoid illegitimate offspring. Proc. Natl. Acad. Sci. 91:1992-1997.

Matton, D. P., Mau, S. L., Okamoto, S., Clarke, A. E., and Newbigin, E. 1995. The S-locus of *Nicotiana alata*: genomic organization and sequence analysis of two S-RNase alleles. Plant Mol. Biol. 28:847-858.

Matton, D. P., Maes, O., Laublin, G., *et al.* 1997. Hypervariable domains of self-incompatibility RNases mediate allele-specific pollen recognition. Plant Cell 9:1757-1766.

Matton, D. P., Luu, D. T., Xike, Q., Laublin, G., O' Brien, M., Maes, O., Morse, D., and Cappadocia, M. 1999. Production of an S RNase with dual specificity suggests a novel hypothesis for the generation of new S alleles. Plant Cell. 11:2087-2097.

Maruyama, K., and Tachida, H. 1992. Genetic variability and geographic structure in partially selfing populations. Japanese J. of Genetics 67:39-51.

Mau, S. L., Williams, E. G., Atkinson, A., Anderson, M. A., Cornish, E. C., Greco, B., Simpson, R. J., Kheyr-Pour, A., and Clark, A. E. 1986. Style proteins of wild tomato (*Lycopersicon peruvianum*) associated with expression of self-incompatibility. Planta 169:184-191.

Maynard Smith, J., and Haigh, J. 1974. The hitchhiking effect of a favourable gene. Genet. Res. 23:23-35.

Meyers, B. C., Dickerman, A. W., Michelmore, R. W., Sivaramakrishnan, S., Sobral, B. W., and Young, N. D. 1999. Plant disease resistance genes encode members of an ancient and diverse protein family within the nucleotide-binding superfamily. Plant J. 20:317-332.

Miyashita, N., and Langley, C. H. 1988. Molecular and phenotypic variation of the *white* locus region in *Drosophila melanogaster*. Genetics 120:199-212.

Möller, M., Clokie, M., Cubas, P., and Cronk, Q. C. B. 1999. Integrating molecular and developmental genetics: a Gesneriaceae case study. In: Hollingsworth, P. M., Bateman, R. J., and Gornall, R. J. (eds). Molecular systematics and plant evolution. Taylor and Francis, London, pp. 375-402.

Moore, D. M., and Lewis, H. 1965. The evolution of self-pollination in *Clarkia xantiana*. Evolution 9:104-114.

Moriyama, E. N., and Powell, J. R. 1996. Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. 13:261-277.

Moriyama, E. N., and Powell, J. R. 1997. Codon usage bias and tRNA abundance in *Drosophila*. Mol. Evol. 45:514-523.

Morton, B. R., Gaut, B. S., and Clegg, M. T. 1996. Evolution of alcohol dehydrogenase genes in the palm and grass families. Proc. Natl. Acad. Sci. USA 93:11735-11739.

Muller, J. 1981. Fossil pollen records of extant angiosperms. Bot. Rev. 47:1-142.

Mullner, E. W., and Kuhn, L. C. 1988. A stem-loop in the 3' untranslated region mediates iron-dependent regulation of transferrin receptor mRNA stability in the cytoplasm. Cell 53:815-825.

Munte, A., Aguadé, M., and Segarra, C. 1997. Divergence of the *yellow* gene between *Drosophila melanogaster* and *D. subobscura*: recombination rate, codon bias and synonymous substitutions. Genetics 147:165-175.

Nacken, W. K., Huijser, P., Beltran, J. P., Saedler, H., and Sommer, H. 1991. Molecular characterization of two stamen-specific genes,*tap1* and *fil1*, that are expressed in the wild type, but not in the deficiens mutant of *Antirrhinum majus*. Mol. Gen. Genet. 229:129-136.

Nagylaki, T. 1983. The evolution of multigene families under intrachromosomal gene conversion. Genetics 106:529-548.

Nagylaki, T., and Petes, T. D. 1982. Intrachromosomal gene conversion and the maintenance of sequence homogeneity among repeated genes. Genetics 100: 315-337.

Nei, M. 1987. Molecular evolutionary genetics. Columbia University Press, New York.

Newbigin, E., Anderson, M. A., and Clark, A. E. 1993. Gametophytic self-incompatibility systems. Plant Cell 5:1315-1324.

Nordborg, M. 1997. Structured coalescent processes on different time scales. Genetics 146:1501-1514.

Norioka, N., Norioka, S., Ohnishi, Y., Ishimizu, T., Oneyama, C., Nakanishi, T., and Sakiyama, F. 1996. Molecular cloning and nucleotide sequences of cDNAs encoding S-allele specific stylar RNases in a self-incompatible cultivar and its self-compatible mutant of Japanese pear, *Pyrus pyrifolia* Nakai. J. Biochem. 120:335-345.

Oberholzer, V., Durbin, M. L., and Clegg M. T. 2000. Comparative genomics of chalcone synthase and *Myb* genes in the grass family. Genes Genet. Syst. 75: 1-16.

Olmstead, R. G., and Reeves, P. 1995. Evidence for the polyphyly of the Scrophulariaceae based on chloroplast *rbc*L and *ndh*F sequences. Ann. Missouri Bot. Gard. 82: 176-193.

Ohgi, K., Horiuchi, H., Watanabe, H., Iwama, M., Takagi, M., and Irie, M. 1992. Evidence that three histidine residues of a base non-specific and adenylic acid preferential ribonuclease from *Rhyzopus niveus* are involved in the catalytic function. J. Biochem. 112:132-138.

Ohta, T. 1981. Genetic variation in small multigene families. Gent. Res. 37:133-149.

Ohta, T. 1984. Some models of gene conversion for treating the evolution of multigene families. Genetics 106:517-528.

Ohta, T. 1992. The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. 23:263-286.

Oxley, D., Munro, S. L., Craik, D. J., and Bacic, A. 1998. Structure and distribution of N-glycans on the S7-allele stylar self-incompatibility ribonuclease of *Nicotiana alata*. J. Biochem. 123:978-983.

Pan, Q., Liu, Y.-S., Budai-Hadriana, O., Sela, M., Carmel-Gorenc, L., Zamirc, D., and Fluhra, R. 2000. Comparative genetics of nucleotide binding site-leucine rich repeat resistance gene homologues in the genomes of two dicotyledons: Tomato and *Arabidopsis*. Genetics 155:309-322.

Pandey, K. K. 1957. Genetics of incompatibility in *Physalis ixocarpa* Brot. A new system. Am. J. Botan. 44:879-887.

Pandey, N. B., Williams, A. S., Sun, J. H., Brown, V. D., Bond, U., and Marzluff, W. F. 1994. Point mutations in the stem-loop at the 3' end of mouse histone mRNA reduce expression by reducing the efficiency of 3' end formation. Mol. Cell Biol. 14:1709-1720.

Parry, S., Newbigin, E., Craik, D., Nakamura, K. T., Bacic, A., and Oxley, D. 1998. Structural analysis and molecular model of a self-incompatibility RNase from wild tomato. Plant Physiol. 116:463-469.

Parsch, J., Braverman, J. M., and Stephan, W. 2000. Comparative sequence analysis and patterns of covariation in RNA secondary structures. Genetics 154:909-921.

Pollak, E. 1987. On the theory of partially inbreeding finite populations. I. Partial selfing. Genetics 117:353-360.

Powell, J. R. 1997. Progress and prospects in evolutionary biology. The *Drosophila* model. Oxford University Press, Oxford.

Purugganan, M. D. 2000. The molecular population genetics of regulatory genes. Mol. Ecol. 9:1451-1461.

Rabinowicz, P. D., Braun, E. L., Wolfe, A. D., Bowen, B., and Grotewold, E. 1999. Maize *R2R3 Myb* genes: Sequence analysis reveals amplification in the higher plants. Genetics 153:427-444.

Reeves, P. A., and Olmstead, R. G. 1998. Evolution of novel morphological and reproductive traits in a clade containing *Antirrhinum majus* (Scrophulariaceae). American J. Bot. 85:1047-1056.

Richards, A. J. 1997. Plant breeding systems. Chapman and Hall, New York.

Richman A. D., Broothaerts W., and Kohn J. R. 1997. Self-incompatibility RNases from three plant families: homology or convergence? A. J. Botany 84: 912-917.

Richman, A. D., Kao, T. H., Schaeffer, S.W., and Uyenoyama, M. K. 1995. S-allele sequence diversity in natural populations of *Solanum carolinense* (Horsenettle). Heredity 75:405-415.

Richman, A. D., Uyenoyama, M. K., and Kohn, J. R. 1996a. S-allele diversity in a natural population of *Physalis crassifolia* (Solanaceae) (ground cherry) assessed by RT-PCR. Heredity 76:497-505.

Richman, A. D., Uyenoyama, M. K., and Kohn, J. R. 1996b. Allelic diversity and gene genealogy at the self-incompatibility locus in the Solanaceae. Science 273:1212-1216.

Richman, A. D., and Kohn, J. R. 1999. Self-incompatibility alleles from *Physalis*: implications for historical inference from balanced genetic polymorphisms. Proc. Natl. Acad. Sci. U S A 96:168-172.

Richman, A. D., and Kohn, J. R. 2000. Evolutionary genetics of self-incompatibility in the Solanaceae. Plant Mol. Biol. 42:169-179.

Rick, C. M., Forbes, J. F., and Holle, M. 1977. Genetic variation in *Lycopersicon pimpinellifolium*: evidence of evolutionary change in mating systems. Pl. Evol. 127:139-170.

Rick, C. M., Forbes, J. F., and Tanksley, S. D. 1979. Evolution of mating systems in *Lycopersicon hirsutum* as deduced from genetic variation in electrophoretic and morphological characteres. Pl. Syst. Evol. 132:279-298.

Ritland, K. 1989. Genetic differentiation, diversity and inbreeding in the mountain monkeyflower (*Mimulus caespitosus*) of the Washington Cascades. Canad. J. Bot. 67:2017-2024.

Ritland, C., and Ritland, K. 1989. Variation of sex allocation among eight taxa of the*Mimulus guttatus* species complex (Scrophulariaceae). Am. J. Bot. 76:1731-1739.

Rothmaler, W. 1956. Taxonomische monographie der gattung *Antirrhinum*. Feddes Repert 136:1-134.

Rozas, J., and Rozas, R. 1997. DnaSP version 2.0: a novel software package for extensive molecular population genetics analysis. Comput. Applic. Biosci. (CABIOS) 13:307-311.

Saba-El-Leil, M., Rivard, S., Morse, D., and Cappadocia, M. 1994. The S11 and S13 self incompatibility alleles in *Solanum chacoense* Bitt. are remarkably similar. Plant Mol. Biol. 24:571-583.

Sanmiguel, P., Tikhonov, A., Jin, Y. K., Motchoulskaia, N., Zakharov, D., Melakeberhan, A., Springer, P. S., Edwards, K. J., Lee, M., Avramova, Z., and Bennetzen, J. L. 1997. Nested retrotransposons in the intergenic regions of the maize genome. Science 274:765-768.

Sassa, H., Nishio, T., Kowyama, Y., Hirano, H., Koba, T., and Ikehashi, H. 1996. Self-incompatibility (S) alleles of the Rosaceae encode members of a distinct class of the T2/S ribonuclease superfamily. Mol. Gen. Genet. 250:547-557.

Savolainen, O., Langley, C. H., Lazzaro, B. P., and Fréville, H. 2000. Contrasting patterns of nucleotide polymorphism at the alcohol dehydrogenase locus in the outcrossing *Arabidopsis lyrata* and the selfing *Arabidopsis thaliana*. Mol. Biol. Evol. 17:645-655.

Schemske, D. W., and Lande, R. 1985. The evolution of self fertilization and inbreeding depression in plants. II. Empirical observations. Evolution 39:41-52.

Schoen, D. J. 1982. The breeding system of *Gilia achilleifolia*: variation in floral characteristics and outcrossing rate. Evolution 36:352-360.

Schoen, D. J., and Brown, A. H. D. 1991. Intraspecific variation in population gene diversity and effective population size correlates with the mating system in plants. Proc. Natl. Acad. Sci. USA 88:4494-4497.

Schopfer, C. R., Kochs, G., Lottspeich, F., and Ebel, J. 1998. Molecular characterization and functional expression of dihydroxypterocarpan 6a-hydroxylase, an enzyme specific for pterocarpanoid phytoalexin biosynthesis in soybean (*Glycine max* L.). FEBS Lett. 432:182-186.

Schwarz-Sommer, Z.., Hue, I., Huijser, P., Flor, P. J., Hansen, R., Tetens, F., Lonnig, W. E., Saedler, H., and Sommer, H. 1992. Characterization of the *Antirrhinum* floral homeotic MADS-box gene *deficiens*: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. EMBO 11:251-263.

Schwarz-Sommer, Z., Nacken, W., Saedler, H., and Sommer, H. 1990. Genetic control of flower development by homeotic genes in *Antirrhinum majus.* Science 250:931-936.

Sharma, N., and Shivanna, K. R. 1983. Lectin-like components of pollen and complementary saccharide moiety of the pistil are involved in self-incompatibility recognition. Curr. Sci. 52:913-916.

Sharma, N., Bajaj, M., and Shivanna, K. R. 1985. Overcomming self-incompatibility through the use of lectins and sugars in *Petunia* and *Eruca*. Ann. Bot. 55:139-141.

Sharp, P. M., Stenico, M., Peden, J. F., and Lloyd, A. T. 1993. Codon usage: mutational bias, translational selection, or both? Biochem. Soc. Trans. 21:835-841

Sherman, M. 1939. The inheritance of self-sterility in certain species of *Antirrhinum*. Z. Indukt. Abstammungs-Vererbungsl. 77:1-17.

Shinozuka, Y., Kojima, S., Shomura, A., Ichimura, H., Yano, M., Yamamoto, K., and Sasaki,T. 1999. Isolation and characterization of rice MADS box gene homologues and their RFLP mapping. DNA Res. 6:123-129.

Sims, T. L. 1993. Genetic regulation of self-incompatibility. Critical Reviews in Plant Sciences12:129-167.

Small, R. L., Ryburn, J. A., and Wendel, J. F. 1999. Low levels of nucleotide diversity at homoeologous *Adh* loci in allotretraploid cotton (*Gossypium* L.). Mol. Biol. Evol. 16:491-501.

Small, R. L., and Wendel, J. F. 2000. Copy number lability and evolutionary dynamics of the *Adh* gene family in diploid and tetraploid cotton. Genetics 155:1913-1926.

Solbrig, O. T., and Rollins, R. C. 1977. The evolution of autogamy in species of the mustard genus *Leavenworthia*. Evolution 31:265-281.

Soltis, P. S., and Soltis, D. E. 2000. Contributions of plant molecular systematics to studies of molecular evolution. Plant Mol. Biol. 42:45-75.

Soltis, P. S., Soltis, D. E., and Chase, M. W. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature 402:402-404.

Soltis, D. E., Soltis, P. S., Nickrent, D. L., *et al.* 1997. Angiosperm phylogeny inferred from 18S ribosomal DNA sequences. Ann. Missouri Bot. Gard. 84:1-49.

Sommer, H., Carpenter, R., Harrison, B. J., and Seadler, H. 1985. The transposable element TAM3 of *Antirrhinum majus* generates a novel type of sequence alterations upon excision. Mol. Gen. Genet. 199:225-231.

Sommer, H., and Seadler, H. 1986. Structure of the chalcose synthase gene of *Antirrhinum majus*. Mol. Gen. Genet. 202:429-434.

Stebbins, G. L. 1957. Self-fertilization and population variation in the higher plants. American Naturalist 91:337-354.

Steinmayr, M., Motte, P., Sommer, H., Saedler, H., and Schwarz-Sommer, Z. 1994. FIL2, an extracellular Leucine-Rich Repeat protein, is specifically expressed in *Antirrhinum* flowers. Plant J. 5:459-467.

Stephan, W., and Langley, C. H. 1998. DNA polymorphism in *Lycopersicon* and crossing-over per physical length. Genetics 150:1585-1593.

Sutliff, T. D., Huang, N., Litts, J. C., and Rodriguez, R. L. 1991. Characterization of an alpha-amylase multigene cluster in rice. Plant Mol. Biol. 16:579-591.

Sutton, D. A. 1988. A revision of the tribe Antirrhineae. Oxford Univ Press, London

Tanaka, T., and Nei, M. 1989. Positive darwinian selection observed at the variable-region genes of immunoglobulins. Mol Biol. Evol. 6:447-459.

Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.

Tajima, F. 1993. Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135:599-607.

Tao, R., Yamane, H., and Sugiura, A. 1999. Cloning of genomic DNA sequences encoding S1, S3, S4, and S6-RNases (Accession Nos. AB03185, AB03186, AB03187 and AB01388) from sweet cherry (*Prunus avium* L.) (PGR99-166) Plant Physiol. 121:1057.

Taylor, C. B., Bariola, P. A., delCardayre, S. B., Raines, R. T., and Green, P. J. 1993. RNS2: a senescence-associated RNase of *Arabidopsis* that diverged from the S-RNases before speciation. Proc. Natl. Acad. Sci. USA. 90:5118-5122.

Terryn, N, Heijnen L, De Keyser A, *et al.* 1999. Evidence for an ancient chromosomal duplication in *Arabidopsis thaliana* by sequencing and analyzing a 400-kb contig at the APETALA2 locus on chromosome 4. FEBS Lett. 445:237-245.

Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J. T., Munster, T., Winter, K. U., and Saedler, H. 2000. A short history of MADS-box genes in plants. Plant Mol. Biol. 42:115-149.

Thompson, J., Gibson, T. J., Plewniak, F., Jeanmougin, F., and Higgins, D. G. 1997. The ClustalX window interface: flexible stategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. 25:4876-4882.

Todokoro, S., Terauchi, R., and Kawano, S. 1995. Microsatellite polymorphisms in natural populations of *Arabidopsis thaliana* in Japan. Japanese J. Genetics 70:543-554.

Treuren, R. V., Kuittinen, H., Karkkainen, K., Baena-Gonzalez, E., and Savolainen, O. 1997. Evolution of microsatellites in *Arabis petraea* and *Arabis lyrata*, outcrossing relatives of *Arabidopsis thaliana*. Mol. Biol. Evol. 14:220-229.

Tröbner, W., Ramirez, L., Motte, P., Hue, I., Huijser, P., Lonnig, W. E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. 1992. GLOBOSA: a homeotic gene which interacts with DEFICIENS in the control of *Antirrhinum* floral organogenesis. EMBO J. 11:4693-4704.

Tsai, D. S., Lee H. S., Post, L. C., Kreiling, K. M., and Kao, T. H. 1992. Sequence of an S-protein of *Lycopersicon peruvianum* and comparison with other solanaceous S-proteins. Sex. Plant Reprod. 5:256-263.

Tseng, H. 1938. Self-sterility in *Antirrhinum* and *Petunia*. J. Genet. 36:127-138.

Uyenoyama, M. K. 1997. Genealogical structure among alleles regulating self-incompatibility in natural populations of flowering plants. Genetics 147:1389-400.

Upadhyay, K. C., Sommer, H., Krebbers, E., and Seadler, H. 1985. The paramutagenic line *niv*-44 has a 5 kb insert, TAM2, in the chalcone synthase gene of *Antirrhinum majus*. Mol. Gen. Genet. 199:201-207.

Ushijima, K., Sassa, H., Tao, R., Yamane, H., Dandekar, A. M., Gradziel, T. M., and Hirano, H. 1998. Cloning and characterization of cDNAs encoding S-RNases from almond (*Prunus dulcis*): primary structural features and sequence diversity of the S-RNases in Rosaceae. Mol. Gen. Genet. 260:261-268.

Vasek, F. C., and Harding, J. 1976. Outcrossing in natural populations. IV. Analysis of outcrossing, inbreeding, and selection in *Clarkia tembloriensis*. Evolution 30:403-411.

Vekemans, X., and Slatkin, M. 1994. Gene and allelic genealogies at a gametophytic self-incompatibility locus. Genetics 137:1157-1165.

Vieira, C. P., and Charlesworth, D. 2000. Low diversity and divergence in the *fil1* gene family of *Antirrhinum* (Scrophulariaceae). J. Mol. Evol., in press

Vieira, C. P., Vieira, J., and Charlesworth, D. 1999. Evolution of the cycloidea gene family in *Antirrhinum* and *Misopates*. Mol. Biol. Evol. 16:1474-1483.

Wade, M. J., and McCauley, D. E. 1988. Extinction and re-colonization: their effects on the genetic differentiation of local populations. Evolution 42:995-1005.

Wagstaff, S. J., Heenan, P. B., and Sanderson, M. J. 1999. Classification, origins, and patterns of diversification in New Zealand *Carmichaelinae* (Fabaceae). Am. J. Bot. 86:1346-1356.

Walker, E. A., Ride, J. P., Kurup, S., Franklin-Tong, V. E., Lawrence, M. J. *et al.* 1996. Molecular analysis of two functional homologues of the S3 allele of the *Papaver rhoeas* self-incompatibility gene isolated from different populations. Plant Mol. Biol. 30: 983-994.

Wall, J. D. 1999. Recombination and the power of statistical tests of neutrality. Genet. Res. 74:65-79.

Walsh, B. J. 1985. Interaction of selection and biased gene conversion in a multigene family. Proc. Natl. Acad. Sci. USA 82:153-157.

Wang, T. W., and Arteca, R. N. 1995. Identification and characterization of cDNAs encoding ethylene biosynthetic enzymes from *Pelargonium* x *hortorum* cv Snow Mass leaves. Plant Physiol. 109:627-636.

Wang, R. L., Stec, A., Hey, J., Lukens, L., and Doebley, J. 1999. The limits of selection during maize domestication. Nature 398:236-239.

Weller, S. G., Donoghue, M. J., and Charlesworth, D. 1995. The evolution of self-incompatibility in the flowering plants: a phylogenetic approach. In: Hoch, P. C.,

and Stephenson, A. G. (eds.). Experimental and molecular approaches to Plant Biosystematics. Missouri Botanic Garden, St Louis Mo., pp. 355-382.

Wheldale, M. 1907. The inheritance of flower colour in *Antirrhinum majus*. Proc. R. Soc. Lond. 79B:288-304.

Wolfe, A. D., and dePamphilis, C. W. 1998. The effect of relaxed functional constraints on the photosynthetic gene *rbc*L in photosynthetic and nonphotosynthetic parasitic plants. Mol. Biol. Evol. 15:1243-1258.

Wolfe, K. H., Sharp, P. M., and Li, W. H. 1989. Rates of synonymous substitution in plant nuclear genes. J. Mol. Evol. 29:208-211.

Woodward J.R., Bacic A., Jahnen W., and Clarke A. E. 1989. N-linked glycan chains on S-allele-associated glycoproteins from *Nicotiana alata*. Plant Cell 1:511-514.

Wright, F. 1990. The "effective number of codons" used in a gene. Gene 87:23-29.

Xue, Y., Carpenter, R., Dickinson, H. G., and Coen, E. S. 1996. Origin of allelic diversity in *Antirrhinum* S locus RNase. The Plant Cell 8:805-814.

Yamaguchi-Shinozaki, K., Mundy, J., and Chua, N. H. 1990. Four tightly linked rab genes are differentially expressed in rice. Plant Mol. Biol. 14:29-39.

Ye, Z. H., and Droste, D. L. 1996. Isolation and characterization of cDNAs encoding xylogenesis-associated and wounding-induced ribonucleases in *Zinnia elegans*. Plant Mol. Biol. 30:697-709.

Yu, D., Kotilainen, M., Pollanen, E., Mehto, M., Elomaa, P., Helariutta, Y., Albert,V. A., and Teeri, T. H. 1999. Organ identity genes and modified patterns of flower development in *Gerbera hybrida*. Plant J. 17:51-62.

Zimmer, E. A., Martin, S. L., Beverley, S. M., Kan, Y. W., and Wilson, A. C. 1980. Rapid duplication and loss of genes coding for the alpha chains of hemoglobin. Proc. Natl. Acad. Sci. USA 77:2158-2162.

Zuckerkandl, E., and Pauling, L. 1965. Evolutionary divergence and convergence in proteins. In: Bryson, B., and Vogel, H. J. (eds.). Evolving genes and proteins. Academic Press, New York, pp. 97-116.

Zurek, D. M., Mou, B., Beecher, B., and McClure, B. 1997. Exchanging sequence domains between S-RNases from *Nicotiana alata* disrupts pollen recognition. Plant J. 11:797-808.

Appendix

Phylogenetic relationships of the second clade of traditional Asteridea, according to Chase *et al.* 1993.